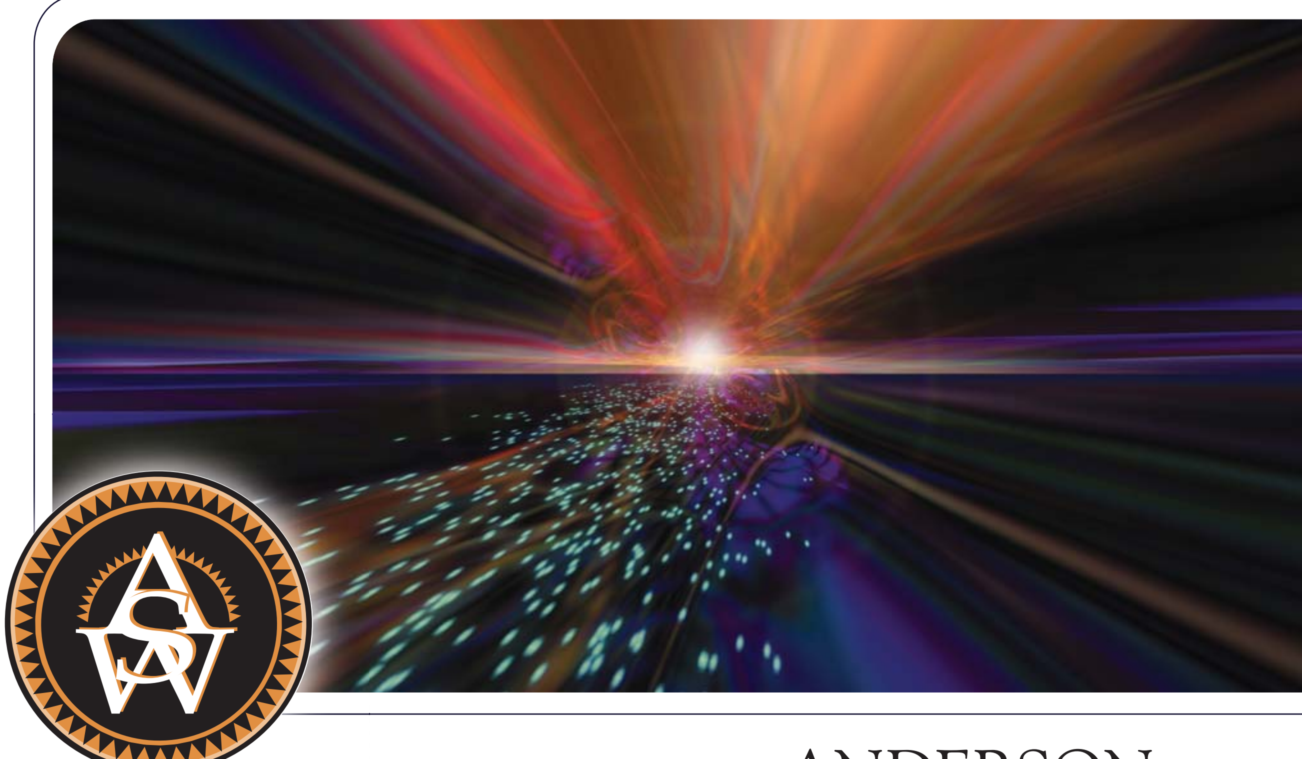


10a. edición

ESTADÍSTICA PARA ADMINISTRACIÓN Y ECONOMÍA



ANDERSON
SWEENEY
WILLIAMS



Estadística para administración y economía

10a. edición

David R. Anderson
University of Cincinnati

Dennis J. Sweeney
University of Cincinnati

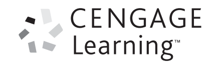
Thomas A. Williams
Rochester Institute of Technology

Traducción:
Ma. del Carmen Hano Roa
Diplom Mathematekirin
Ludwig-Maximiliams Universität München, Alemania

Revisión técnica:
Dra. Teresa López Álvarez
Consultora independiente

CENGAGE Learning™

Australia • Brasil • Corea • España • Estados Unidos • Japón • México • Reino Unido • Singapur



***Estadística para administración y economía*, 10a. edición**
Anderson, David R., Dennis J. Sweeney y Thomas A. Williams

Presidente de Cengage Learning Latinoamérica:
Javier Arellano Gutiérrez

Director General México y Centroamérica:
Héctor Enrique Galindo Iturribarría

Director Editorial Latinoamérica:
José Tomás Pérez Bonilla

Editor:
Sergio R. Cervantes González

Director de producción:
Raúl D. Zendejas Espejel

Editor de producción:
Timoteo Eliosa García

Ilustrador:
Michael Stratton/cmiller design

Diseño de portada:
Paul Neff

Imagen de portada:
© Brand X Images/Getty Images

Composición tipográfica:
José Jaime Gutiérrez Aceves

Traducido del libro *Statistics for Business and Economics*, 10th ed.
Publicado en inglés por Thomson/Southwestern © 2008
ISBN: 0-324-36068-1
Datos para catalogación bibliográfica:
Anderson, David R., Dennis J. Sweeney y Thomas A. Williams
Estadística para administración y economía, 10a. ed.
ISBN-13: 978-607-481-319-7
ISBN-10: 607-481-319-1

Visite nuestro sitio en:
http://latinoamerica.cengage.com

Dedicado a
Marcia, Cherri y Robbie

Contenido breve

Contenido

Capítulo 5 Distribuciones de probabilidad discreta 186

Capítulo 6 Distribuciones de probabilidad continua 225

Capítulo 7 Muestreo y distribuciones muestrales 257

Capítulo 8 Estimación por intervalo 299

Capítulo 9 Prueba de hipótesis 338

Capítulo 11 Inferencias acerca de varianzas poblacionales 434

Capítulo 12 Pruebas de bondad de ajuste e independencia 457

Capítulo 13 Diseño de experimentos y análisis de varianza 490

Capítulo 14 Regresión lineal simple 543

Capítulo 22 Encuestas muestrales 915

Prefacio

El propósito de *Estadística para administración y economía* es proporcionar, en especial a los estudiantes de las áreas de la administración y de la economía, una introducción conceptual al campo de la estadística y de sus aplicaciones. El texto está orientado a las aplicaciones y ha sido escrito pensando en las necesidades de quienes no son matemáticos; los conocimientos matemáticos requeridos son los conocimientos del álgebra.

Las aplicaciones del análisis de datos y de la metodología estadística son parte integral de la presentación y organización del material de este libro. El estudio y el desarrollo de cada técnica se presentan mediante una aplicación, en donde los resultados estadísticos permiten entender las decisiones y la solución del problema presentado.

Aunque el libro está orientado hacia las aplicaciones, hemos tenido cuidado de presentar un desarrollo metodológico sólido y de emplear la notación convencional al tópico que se estudia. De esta manera, los estudiantes encontrarán que este libro les proporciona una buena preparación para el estudio de material estadístico más avanzado. En el apéndice A se proporciona una bibliografía que servirá como guía para un estudio más profundo.

El libro introduce al estudiante a los paquetes de software Minitab de Microsoft y a Excel haciendo énfasis en el papel que tiene el software en la aplicación del análisis estadístico. Minitab se presenta como uno de los principales paquetes de software para estadística, tanto en la enseñanza, como en la práctica. Excel no es un paquete de software para estadística, pero su amplia disponibilidad y uso lo hacen relevante para que los estudiantes conozcan las posibilidades de Excel para la estadística. El empleo de Excel y Minitab se presenta en los apéndices, permitiendo así al profesor la suficiente flexibilidad para dar tanta importancia al uso de la computadora como él lo desee.

Cambios en la 10a. edición

Agradecemos la acogida y la respuesta positiva a las ediciones anteriores de *Estadística para administración y economía*. Por tanto, al hacer modificaciones en esta nueva edición, hemos conservado el mismo estilo de presentación y la sencillez de esas ediciones. Los cambios más importantes hechos en esta nueva edición se presentan a continuación.

Cambios al contenido

En seguida se resumen algunos de los cambios que hemos hecho al contenido en esta edición.

- **Valores-*p*** En la edición anterior insistimos en el uso de los valores-*p* en las pruebas de hipótesis. En esta edición hacemos lo mismo, no obstante, hemos hecho más sencilla la introducción a los valores-*p* simplificando la definición conceptual. Ahora dice: "Un valor-*p* es una probabilidad que mide la evidencia contra la hipótesis nula que proporciona la muestra. Entre menor es el valor-*p*, mayor es la evidencia contra H_0." Después de esta definición conceptual, se presentan las definiciones operacionales que explican cómo calcular el valor-*p* en pruebas de la cola izquierda (cola inferior), de la cola derecha (cola superior) y de dos colas. Con la experiencia hemos aprendido que el separar la definición conceptual de las definiciones operacionales ayuda al estudiante a entender con más facilidad el nuevo material.
- **Procedimientos de Minitab y de Excel para calcular el valor-*p*.** Algo nuevo en esta edición es un apéndice en el que se demuestra cómo se usan Minitab y Excel para calcular valores-*p* relacionados con los estadísticos de prueba z, t, χ^2 y F. A los estudiantes que emplean una calculadora manual para calcular los estadísticos de prueba se les enseña có-

mo usar las tablas estadísticas para dar un intervalo de valores-*p*. En el apéndice F se les explica la forma de calcular con exactitud el valor-*p* usando Minitab o Excel. Este apéndice es de utilidad al estudiar las pruebas de hipótesis en los capítulos 9 a 16.

- **Tabla de la distribución normal estándar acumulada.** A muchos de nuestros usuarios puede sorprenderles que en esta nueva edición usemos tablas de distribución normal estándar acumulada. Hemos hecho este cambio porque creemos que la tendencia es que cada vez más estudiantes y profesionistas hagan uso del software para computadoras. Antes, todo mundo empleaba las tablas porque era la única fuente de información acerca de la distribución normal. Sin embargo, hoy muchos estudiantes están dispuestos a aprender a usar el software para estadística. Los estudiantes encontrarán que casi todos los paquetes de software usan la distribución normal estándar acumulada. Por tanto, es cada vez más importante que en un libro de introducción a la estadística se usen las tablas de probabilidad normal que el estudiante encontrará cuando trabaje con el software para estadística. No es deseable usar un tipo de tablas para la distribución normal estándar en el libro y otro tipo diferente cuando se usen los paquetes de software. Aquellas personas que usen por primera vez la tabla de distribución normal acumulada encontrarán que, en general, estas tablas facilitan los cálculos de la distribución normal. En particular, una tabla de probabilidad normal acumulada facilita el cálculo de los valores-*p* en las pruebas de hipótesis.
- **Diseño de experimentos y análisis de varianza.** El capítulo 13 se ha reducido y ahora comienza con una introducción a los conceptos del diseño de experimentos. Se tratan también el diseño completamente aleatorizado, el diseño de bloque aleatorizado y los experimentos factoriales. El análisis de varianza se presenta como la técnica fundamental para el análisis de estos diseños. También mostramos que el procedimiento de análisis de varianza puede emplearse en estudios observacionales.
- **Otras modificaciones al contenido.** Las siguientes adiciones se encontrarán en la nueva edición:

  – En el capítulo 1 se presentan ejemplos nuevos de datos de series de tiempo.
  – En el capítulo 2 el apéndice sobre Excel ahora proporciona instrucciones más completas acerca de cómo elaborar una distribución de frecuencia y un histograma con datos cuantitativos.
  – Revisamos los lineamientos acerca del tamaño de la muestra necesario para el uso de de la distribución *t*, lo que es consistente con el uso de la distribución *t* en los capítulos 8, 9 y 10.
  – El capítulo 17 ha sido actualizado con números índices de uso corriente.
  – Ahora en el manual de soluciones se encuentran los pasos para la solución de los ejercicios usando la distribución normal acumulada y más detalles en las explicaciones de cómo calcular los valores-*p* en las pruebas de hipótesis.

Ejemplos y ejercicios nuevos a partir de datos reales

Hemos agregado 200 ejemplos y ejercicios nuevos con base en datos reales y en fuentes de referencias recientes sobre información estadística. Con datos obtenidos de fuentes empleadas también por *Wall Street Journal*, *USA Today*, *Fortune*, *Barron's* y otras, hemos empleado estudios actuales para elaborar explicaciones y crear ejercicios que demuestren los diversos usos de la estadística en la administración y la economía. Pensamos que el uso de datos reales generará más interés en los estudiantes por este material y les permitirá aprender más acerca de la metodología estadística y de sus aplicaciones. Esta 10a. edición contiene 350 ejemplos y ejercicios basados en datos reales.

Casos problema nuevos

En esta edición hemos agregado seis casos problema nuevos, con lo que la cantidad de casos problema en este libro se eleva a 31. Los casos problema nuevos aparecen en los capítulos sobre es-

tadística descriptiva, estimación por intervalo y regresión. Estos casos problema proporcionan a los estudiantes la oportunidad de analizar conjuntos de datos un poco mayores y de elaborar reportes administrativos basados en los resultados del análisis.

Características y pedagogía

Los autores Anderson, Sweeney y Williams han conservado en esta edición muchas de las características de las ediciones previas. Las más importantes para los estudiantes se anotan a continuación.

La estadística en la práctica

Cada capítulo empieza con un artículo sobre la estadística en la práctica que describe una aplicación de la metodología estadística que se estudiará en el capítulo. En esta edición los artículos sobre estadística en la práctica de Duke Energy, Rohm and Hass Company y la Food and Drug Administration de Estados Unidos son nuevos.

Ejercicios sobre los métodos y ejercicios de aplicación

Los ejercicios al final de cada sección se dividen en dos partes, métodos y aplicaciones. Los ejercicios sobre los métodos requieren del estudiante el uso de las fórmulas para hacer los cálculos necesarios. Los ejercicios de aplicación demandan que el estudiante use el material del capítulo en una situación de la vida real. De esta manera, los estudiantes dan atención, primero, a los cálculos y después a las sutilezas de la aplicación e interpretación de la estadística.

Ejercicios de autoexamen

Algunos ejercicios son ejercicios de autoexamen. Las soluciones completas de estos ejercicios se proporcionan en el apéndice D, al final del libro. Los estudiantes pueden hacer estos ejercicios de autoexamen y verificar de inmediato la solución para evaluar su comprensión de los conceptos presentados en el capítulo.

Anotaciones al margen, notas y comentarios

Anotaciones al margen que resaltan puntos clave y proporcionan una explicación adicional para el estudiante son características esenciales de este libro. Estas anotaciones, que aparecen al margen, tienen el propósito de enfatizar y mejorar la comprensión de los términos y conceptos que se presentan en el texto.

Al final de cada sección, presentamos notas y comentarios que tienen por objeto aclarar aún más la metodología estadística y su aplicación. Las notas y los comentarios contienen advertencias sobre la metodología o limitaciones de ésta, recomendaciones para su aplicación, breves descripciones de otras consideraciones técnicas y otros asuntos.

Archivos de datos que vienen con el texto

En el disco compacto que viene con el libro se encuentran más de 200 archivos de datos. Estos archivos vienen tanto en formato para Minitab como para Excel. En el texto se usan logotipos para indicar conjuntos de datos disponibles en el disco compacto. También hay conjuntos de datos para los casos problema, así como conjuntos de datos para ejercicios más grandes.

Material de apoyo para el profesor

Este libro cuenta con una serie de recursos para el profesor, los cuales están disponibles en inglés y sólo se proporcionan a los docentes que lo adopten como texto en sus cursos.

Para direcciones de correo electrónico:

Cengage Learning México y Centroamérica	clientes.mexicoca@cengage.com
Cengage Learning Caribe	clientes.caribe@cengage.com
Cengage Learning Cono Sur	clientes.conosur@cengage.com
Paraninfo	clientes.paraninfo@cengage.com
Colombia	clientes.pactoandino@cengage.com

Además encontrará más apoyos en el sitio web de este libro:

http://latinoamerica.cengage.com/anderson

Las direcciones de los sitios web referidas a lo largo del texto no son administradas por Cengage Learning Latinoamérica, por lo que ésta no es responsable de los cambios para mantenerse al tanto de cualquier actualización.

Agradecimientos

Un agradecimiento especial a nuestros colegas de las empresas y de la industria que nos proporcionaron el material para *Estadística para administración y economía*. A cada uno le damos un reconocimiento individual en la línea de créditos que aparece en cada uno de los artículos. Por último agradecemos a nuestros editores, Charles McCormick, Jr. y Alice Denny, a nuestro administrador de proyecto, Amy Hackett, a nuestro director de mercadotecnia, Larry Qualls, y a todos los colaboradores de Thomson South-Western por su asesoría y apoyo editorial durante la elaboración de este libro.

David R. Anderson
Dennis J. Sweeney
Thomas A. Williams

Acerca de los autores

David R. Anderson. Profesor de análisis cuantitativo en el College of Business Administration de la Universidad de Cincinnati. Nació en Grand Forks, Dakota del Norte, y obtuvo los grados académicos B.S., M.S. y Ph.D. en la Purdue University. El profesor Anderson ha sido director del Department of Quantitative Analysis and Operations y decano asociado de la College of Business Administration. Además, fue coordinador del primer Executive Program de la escuela.

En la Universidad de Cincinnati, el profesor Anderson ha dado cursos introductorios de estadística para estudiantes de administración, así como cursos a nivel de posgrado sobre análisis de regresión, análisis multivariado y ciencia de la administración. También ha impartido cursos de estadística en el Departamento del Trabajo en Washington, D. C. Ha sido honrado con nominaciones y premios de excelencia en la enseñanza y en la atención a organizaciones estudiantiles.

El profesor Anderson es coautor de diez libros en las áreas de estadística, ciencias de la administración, programación lineal y producción y administración de operaciones. Es asesor activo en los temas de muestreo y de métodos estadísticos.

Dennis J. Sweeney. Dennis J. Sweeney es profesor de análisis cuantitativo y fundador del Center for Productivity Improvement en la Universidad de Cincinnati. Nació en Des Moines, Iowa, y obtuvo el grado B.S.B.A. en la Drake University y los grados M.B.A. y D.B.A. en la Universidad de Indiana. De 1978 a 1979, el profesor Sweeney trabajó en el grupo de ciencia de la administración de Procter & Gamble; de 1981 a 1982, fue profesor invitado en la Duke University. Ha sido director del Department of Quantitative Analysis y decano asociado de la College of Business Administration en la Universidad de Cincinnati.

El profesor Sweeney ha publicado más de 30 artículos y monografías en las áreas de ciencia de la administración y estadística. Sus investigaciones han sido patrocinadas por The National Science Fundation, IBM, Procter & Gamble, Federated Department Stores, Kroger y Cincinnati Gas & Electric, las cuales han sido publicadas en *Management Science, Operation Research, Mathematical Programming, Decision Sciences* y en otras revistas.

El profesor Sweeney es coautor de diez libros en las áreas de estadística, ciencias de la administración, programación lineal y producción y administración de operaciones.

Thomas A. Williams. Thomas A. Williams es profesor de ciencia de la administración en el College of Business at Rochester Institute of Technology. Nació en Elmira, Nueva York y obtuvo el grado B.S. en la Clarkson University. Realizó su tesis profesional en el Rensselaer Polytechnic Institute, donde obtuvo los grados M.S. y Ph.D.

Antes de integrarse a la College of Business de RIT, el profesor Williams fue miembro de la facultad en el College of Business Administration de la Universidad de Cincinnati, en donde elaboró el programa para Sistemas de la Información, del que fue coordinador. En RIT fue el primer director del Decision Sciences Departament. Imparte cursos de ciencia de la administración y de estadística, así como cursos de análisis de regresión y de decisión.

El profesor Williams es coautor de siete libros en las áreas de estadística, ciencias de la administración, producción y administración de operaciones y matemáticas. Ha sido asesor de múltiples empresas *Fortune 500* y ha trabajado en proyectos que van desde el uso del análisis de datos a la elaboración de modelos de regresión a gran escala.

CAPÍTULO 1



Datos y estadísticas

CONTENIDO



LA ESTADÍSTICA *en* LA PRÁCTICA

*BUSINESSWEEK**
NUEVA YORK, NUEVA YORK

Con una circulación mundial de más de 1 millón de ejemplares, *BusinessWeek* es la revista más leída en el mundo. Más de 200 reporteros y editores especializados en 26 oficinas alrededor del mundo producen diversos artículos de interés para la comunidad interesada en los negocios y la economía. Junto a los artículos principales y los tópicos de actualidad, la revista presenta diversas secciones regulares sobre negocios internacionales, análisis económicos, procesamiento de la información y ciencia y tecnología. La información en las secciones regulares ayuda a los lectores a mantenerse al día de los avances y novedades y a evaluar el impacto de éstos en los negocios y en las condiciones económicas.

La mayor parte de los números de *BusinessWeek* contienen un artículo de fondo sobre algún tema de interés actual. Por ejemplo, el número del 6 de diciembre de 2004 contenía un reportaje especial sobre los precios de los artículos hechos en China; el número del 3 de enero de 2005 proporcionaba información acerca de dónde invertir en 2005 y el número del 4 de abril de 2005 proporcionaba una panorámica de *BusinessWeek 50*, un grupo diverso de empresas de alto desempeño. Además, la revista semanal *BusinessWeek Investor* proporciona artículos sobre el estado de la economía, que comprenden índices de producción, precios de las acciones de fondos mutualistas y tasas de interés.

BusinessWeek también usa métodos e información estadísticos en la administración de su propio negocio. Por ejemplo, una encuesta anual hecha a sus suscriptores le permitió tener datos demográficos sobre sus hábitos de lectura, compras probables, estilo de vida, etc. Los directivos de *BusinessWeek* usan resúmenes estadísticos obtenidos a partir de las encuestas para dar un mejor servicio a sus suscriptores y anunciantes. Mediante una encuesta reciente entre los suscriptores estadounidenses se supo que 90% de los suscriptores de *BusinessWeek* tienen una computadora personal en casa y que 64% de ellos realizan en el trabajo compras por computadora. Estas estadísticas indican a los directivos de *BusinessWeek* que los avances en computación serán de interés para sus suscriptores. Los resultados de la encuesta también le son proporcionados a sus anunciantes potenciales. Los elevados porcentajes de personas que tienen una computadora en casa y que realizan compras por computadora en el trabajo podría ser un incentivo para que los fabricantes de computadoras se anunciaran en *BusinessWeek*.

Este capítulo muestra los tipos de datos con que se cuenta en un análisis estadístico y describe cómo se obtienen los datos. Presenta la estadística descriptiva y la inferencia estadística como medios para convertir los datos en información estadística que tienen un significado y que es fácil de interpretar.



BusinessWeek usa datos y resúmenes estadísticos en muchos de sus artículos. © Terri Millar/E-Visual Communications, Inc.

*Los autores agradecen a Charlene Trentham, Director de investigación de BusinessWeek por proporcionar este artículo para La estadística en la práctica.

Con frecuencia aparece en los periódicos y revistas el siguiente tipo de información:

- La asociación de agentes inmobiliarios informó que la mediana del precio de venta de una casa en Estados Unidos es de $215 000 (*The Wall Street Journal*, 16 de enero de 2006).
- Durante el Super Bowl de 2006 el costo promedio de un spot publicitario de 30 segundos en televisión fue de $2.5 millones (*USA Today*, 27 de enero de 2007).

- En una encuesta de Jupiter Media se encontró que 31% de los hombres adultos ven más de 10 horas de televisión a la semana. Entre las mujeres sólo 26% (*The Wall Street Journal,* 26 de enero de 2004).
- General Motors, uno de los líderes automotrices en descuentos en efectivo da, en promedio, $4300 de incentivo en efectivo por vehículo (*USA Today,* 27 de enero de 2006).
- Más de 40% de los directivos de Marriott Internacional ascienden por escalafón (*Fortune,* 20 de enero de 2003).
- Los Yankees de Nueva York tienen la nómina más alta dentro de la liga mayor de béisbol. En el año 2005 la nómina del equipo fue de $208 306 817, siendo la mediana por jugador de $5 833 334 (*USA Today*, febrero 2006).
- El promedio industrial Dow Jones cerró en 11 577 (*Barron's*, 6 de mayo de 2006).

A los datos numéricos de las frases anteriores se les llama estadísticas. En este sentido el término *estadística* se refiere a datos numéricos, tales como promedios, medianas, porcentajes y números índices que ayudan a entender una gran variedad de negocios y situaciones económicas. Sin embargo, como se verá, el campo de la estadística es mucho más que datos numéricos. En un sentido amplio, la **estadística** se define como el arte y la ciencia de reunir datos, analizarlos, presentarlos e interpretarlos. Especialmente en los negocios y en la economía, la información obtenida al reunir datos, analizarlos, presentarlos e interpretarlos proporciona a directivos, administradores y personas que deben tomar decisiones una mejor comprensión del negocio o entorno económico, permitiéndoles así tomar mejores decisiones con base en mejor información. En este libro se hace hincapié en el uso de la estadística para la toma de decisiones en los negocios y en la economía.

El capítulo 1 empieza con algunos ejemplos de aplicaciones de la estadística en los negocios y en la economía. En la sección 1.2 se define el término *datos* y se introduce el concepto de conjunto de datos. En esta sección se introducen también términos clave como *variables* y *observaciones*, se muestra la diferencia entre datos cualitativos y cuantitativos y se ilustra el uso de datos transversales y de serie de tiempo. En la sección 1.3 se enseña a obtener datos de fuentes ya existentes o mediante encuestas y estudios experimentales diseñados para obtener datos nuevos. Se resalta también el papel tan importante que tiene ahora Internet en la obtención de datos. En las secciones 1.4 y 1.5 se describe el uso de los datos en la estadística descriptiva y para hacer inferencias estadísticas.



Aplicaciones en los negocios y en la economía

En el entorno mundial actual de los negocios y de la economía, todo mundo tiene acceso a enormes cantidades de información estadística. Los directivos y los encargados de tomar decisiones que tienen éxito entienden la información y saben usarla de manera eficiente. En esta sección se proporcionan ejemplos que ilustran algunos de los usos de la estadística en los negocios y en la economía.

Contaduría

Las empresas de contadores públicos al realizar auditorías para sus clientes emplean procedimientos de muestreo estadístico. Por ejemplo, suponga que una empresa de contadores desea determinar si las cantidades en cuentas por cobrar que aparecen en la hoja de balance del cliente representan la verdadera cantidad en cuentas por cobrar. Por lo general, el gran número de cuentas por cobrar hace que su revisión tome demasiado tiempo y sea muy costosa. Lo que se hace en estos casos es que el personal encargado de la auditoría selecciona un subconjunto de las cuentas al que se le llama muestra. Después de revisar la exactitud de las cuentas tomadas en la muestra (muestreadas) los auditores concluyen si la cantidad en cuentas por cobrar que aparece en la hoja de balance del cliente es aceptable.

Finanzas

Los analistas financieros emplean una diversidad de información estadística como guía para sus recomendaciones de inversión. En el caso de acciones, el analista revisa diferentes datos financieros como la relación precio/ganancia y el rendimiento de los dividendos. Al comparar la información sobre una determinada acción con la información sobre el promedio en el mercado de acciones, el analista empieza a obtener conclusiones para saber si una determinada acción está sobre o subvaluada. Por ejemplo, *Barron's* (12 de septiembre de 2005) informa que la relación promedio precio/ganancia de 30 acciones del promedio industrial Dow Jones fue 16.5. La relación precio/ganancia de JPMorgan es 11.8. En este caso la información estadística sobre las relaciones precio/ganancia indican un menor precio en comparación con la ganancia para JPMorgan que el promedio en las acciones Dow Jones. Por tanto el analista financiero concluye que JPMorgan está subvaluada. Ésta y otras informaciones acerca de JPMorgan ayudarán al analista a comprar, vender o a recomendar mantener las acciones.

Marketing

Escáneres electrónicos en las cajas de los comercios minoristas recogen datos para diversas aplicaciones en la investigación de mercado. Por ejemplo, proveedores de datos como ACNielsen e Information Research Inc. compran estos datos a las tiendas de abarrotes, los procesan y luego venden los resúmenes estadísticos a los fabricantes; quienes gastan cientos de miles de dólares por producto para obtener este tipo de datos. Los fabricantes también compran datos y resúmenes estadísticos sobre actividades promocionales como precios o *displays* promocionales. Los administradores de marca revisan estas estadísticas y las propias de las actividades promocionales para analizar la relación entre una actividad promocional y las ventas. Estos análisis suelen resultar útiles para establecer futuras estrategias de marketing para diversos productos.

Producción

La importancia que se le da actualmente a la calidad hace del control de calidad una aplicación importante de la estadística a la producción. Para vigilar el resultado de los procesos de producción se usan diversas gráficas de control estadístico de calidad. En particular, para vigilar los resultados promedio se emplea una gráfica x-barra. Suponga, por ejemplo, que una máquina llena botellas con 12 onzas de algún refresco. Periódicamente un empleado del área de producción toma una muestra de botellas y mide el contenido promedio de refresco. Este promedio o valor x-barra se marca como un punto en una gráfica x-barra. Si este punto queda arriba del límite de control superior de la gráfica, hay un exceso en el llenado, y si queda debajo del límite de control inferior de la gráfica hay falta de llenado. Se dice que el proceso está "bajo control" y puede continuar, siempre que los valores x-barra se encuentren entre los límites de control inferior y superior. Con una interpretación adecuada, una gráfica de x-barra ayuda a determinar si es necesario hacer algún ajuste o corrección a un proceso de producción.

Economía

Los economistas suelen hacer pronósticos acerca del futuro de la economía o sobre algunos aspectos de la misma. Usan una variedad de información estadística para hacer sus pronósticos. Por ejemplo, para pronosticar las tasas de inflación, emplean información estadística sobre indicadores como el índice de precios al consumidor, la tasa de desempleo y la utilización de la capacidad de producción. Estos indicadores estadísticos se utilizan en modelos computarizados de pronósticos que predicen las tasas de inflación.

Aplicaciones de la estadística como las descritas en esta sección integran este libro. Dichos ejemplos proporcionan una visión general de la diversidad de las aplicaciones estadísticas. Como complemento de estos ejemplos, profesionales en los campos de los negocios y de la economía proporcionan los artículos de *La estadística en la práctica* que se encuentran al principio de cada capítulo, en los que se presenta el material que se estudiará en el capítulo. Las aplicaciones en *La estadística en la práctica* muestran su importancia en diversas situaciones de los negocios y la economía.



Datos

Datos son hechos/informaciones y cifras que se recogen, analizan y resumen para su presentación e interpretación. A todos los datos reunidos para un determinado estudio se les llama **conjunto de datos** para el estudio. La tabla 1.1 muestra un conjunto de datos que contiene información sobre 25 empresas que forman parte del S&P 500. El S&P 500 consta de 500 empresas elegidas por Standard & Poor’s. Estas empresas representan 76% de la capitalización de mercado de todas las acciones de Estados Unidos. Las acciones de S&P 500 son estrechamente observadas por los inversionistas y por los analistas de Wall Street.

TABLA 1.1 CONJUNTO DE DATOS DE 25 EMPRESAS S&P 500



Empresa	Bolsa de valores	Denominación abreviada Ticker	Posición en *BusinessWeek*	Precio por acción ($)	Ganancia por acción ($)
Abbott Laboratories	N	ABT	90	46	2.02
Altria Group	N	MO	148	66	4.57
Apollo Group	NQ	APOL	174	74	0.90
Bank of New York	N	BK	305	30	1.85
Bristol-Myers Squibb	N	BMY	346	26	1.21
Cincinnati Financial	NQ	CINF	161	45	2.73
Comcast	NQ	CMCSA	296	32	0.43
Deere	N	DE	36	71	5.77
eBay	NQ	EBAY	19	43	0.57
Federated Dept. Stores	N	FD	353	56	3.86
Hasbro	N	HAS	373	21	0.96
IBM	N	IBM	216	93	4.94
International Paper	N	IP	370	37	0.98
Knight-Ridder	N	KRI	397	66	4.13
Manor Care	N	HCR	285	34	1.90
Medtronic	N	MDT	53	52	1.79
National Semiconductor	N	NSM	155	20	1.03
Novellus Systems	NQ	NVLS	386	30	1.06
Pitney Bowes	N	PBI	339	46	2.05
Pulte Homes	N	PHM	12	78	7.67
SBC Communications	N	SBC	371	24	1.52
St. Paul Travelers	N	STA	264	38	1.53
Teradyne	N	TER	412	15	0.84
UnitedHealth Group	N	UNH	5	91	3.94
Wells Fargo	N	WFC	159	59	4.09

Fuente: Business Week (4 de abril de 2005).

Elementos, variables y observaciones

Elementos son las entidades de las que se obtienen los datos. En el conjunto de datos de la tabla 1.1, cada acción de una empresa es un elemento; los nombres de los elementos aparecen en la primera columna. Como se tienen 25 acciones, el conjunto de datos contiene 25 elementos.

Una **variable** es una característica de los elementos que es de interés. El conjunto de datos de la tabla 1.1 contiene las cinco variables siguientes:

- *Bolsa de valores* (*mercado bursátil*)*:* Dónde se comercializa (cotiza) la acción: N (Bolsa de Nueva York) y NQ (Mercado Nacional Nasdaq).
- *Ticker (denominación abreviada):* Abreviación usada para identificar la acción en la lista de la bolsa
- *Posición en BusinessWeek:* Número del 1 al 500 que indica la fortaleza de la empresa.
- *Precio por acción ($):* El precio de cierre (28 de febrero de 2005).
- *Ganancia por acción ($):* Las ganancias por acción en los últimos 12 meses.

Los valores encontrados para cada variable en cada uno de los elementos constituyen los datos. Al conjunto de mediciones obtenidas para un determinado elemento se le llama **observación**. Volviendo a la tabla 1.1, el conjunto de mediciones para la primera observación (Abbott Laboratories) es N, ABT, 90, 46 y 2.02. El conjunto de mediciones para la segunda observación (Altria Group) es N, MO, 148, 66 y 4.57, etc. Un conjunto de datos que tiene 25 elementos contiene 25 observaciones.

Escalas de medición

La recolección de datos requiere alguna de las escalas de medición siguientes: nominal, ordinal, de intervalo o de razón. La escala de medición determina la cantidad de información contenida en el dato e indica la manera más apropiada de resumir y de analizar estadísticamente los datos.

Cuando el dato de una variable es una etiqueta o un nombre que identifica un atributo de un elemento, se considera que la escala de medición es una **escala nominal**. Por ejemplo, en relación con la tabla 1.1 la escala de medición para la variable bolsa de valores (mercado bursátil) es nominal porque N y NQ son etiquetas que se usan para indicar dónde cotiza la acción de la empresa. Cuando la escala de medición es nominal, se usa un código o una etiqueta no numérica. Por ejemplo, para facilitar la recolección de los datos y para guardarlos en una base de datos en una computadora puede emplearse un código numérico en el que 1 denote la Bolsa de Nueva York y 2 el Mercado Nacional Nasdaq. En este caso los números 1 y 2 son las etiquetas empleadas para identificar dónde cotizan las acciones. La escala de medición es nominal aun cuando los datos aparezcan como valores numéricos.

Una escala de medición para una variable es **ordinal** si los datos muestran las propiedades de los datos nominales y además tiene sentido el orden o jerarquía de los datos. Por ejemplo, una empresa automovilística (Eastside Automotive) envía a sus clientes cuestionarios para obtener información sobre su servicio de reparación. Cada cliente evalúa el servicio de reparación como excelente, bueno o malo. Como los datos obtenidos son las etiquetas excelente, bueno o malo, tienen las propiedades de los datos nominales, pero además pueden ser ordenados o jerarquizados en relación con la calidad del servicio. Un dato excelente indica el mejor servicio, seguido por bueno y, por último, malo. Por lo que la escala de medición es ordinal. Observe que los datos ordinales también son registrados mediante un código numérico. Por ejemplo, en la tabla 1.1 la posición de los datos en *BusinessWeek* es un dato ordinal. Da una jerarquía del 1 al 500 de acuerdo con la evaluación de *BusinessWeek* sobre la fortaleza de la empresa.

Una escala de medición para una variable es una **escala de intervalo** si los datos tienen las características de los datos ordinales y el intervalo entre valores se expresa en términos de una unidad de medición fija. Los datos de intervalo siempre son numéricos. Las calificaciones en una prueba de aptitudes escolares son un ejemplo de datos de intervalo. Por ejemplo, las ca-

lificaciones obtenidas por tres alumnos en la prueba de matemáticas con 620, 550 y 470, pueden ser ordenadas en orden de mejor a peor. Además las diferencias entre las calificaciones tienen significado. Por ejemplo, el estudiante 1 obtuvo $620 - 550 = 70$ puntos más que el estudiante 2 mientras que el estudiante 2 obtuvo $550 - 470 = 80$ puntos más que el estudiante tres.

Una variable tiene una **escala de razón** si los datos tienen todas las propiedades de los datos de intervalo y la proporción entre dos valores tiene significado. Variables como distancia, altura, peso y tiempo usan la escala de razón en la medición. Esta escala requiere que se tenga el valor cero para indicar que en este punto no existe la variable. Por ejemplo, considere el costo de un automóvil. El valor cero para el costo indica que el automóvil no cuesta, que es gratis. Además, si se compara el costo de un automóvil de $30 000, con el costo de otro automóvil, $15 000, la propiedad de razón muestra que $30 000/$15 000 = 2: el primer automóvil cuesta el doble del costo del segundo.

Datos cualitativos y cuantitativos

A los datos cualitativos se les suele llamar datos categóricos.

Los datos también son clasificados en cualitativos y cuantitativos. Los **datos cualitativos** comprenden etiquetas o nombres que se usan para identificar un atributo de cada elemento. Los datos cualitativos emplean la escala nominal o la ordinal y pueden ser numéricos o no. Los **datos cuantitativos** requieren valores numéricos que indiquen cuánto o cuántos. Los datos cuantitativos se obtienen usando las escalas de medición de intervalo o de razón.

El método estadístico adecuado para resumir los datos depende de si los datos son cualitativos o cuantitativos.

Una **variable cualitativa** es una variable con datos cualitativos. El análisis estadístico adecuado para una determinada variable depende de si la variable es cualitativa o cuantitativa. Si la variable es cualitativa, el análisis estadístico es bastante limitado. Tales datos se resumen contando el número de observaciones o calculando la proporción de observaciones en cada categoría cualitativa. Sin embargo, aun cuando para los datos cualitativos se use un código numérico, las operaciones aritméticas de adición, sustracción, multiplicación o división no tienen sentido. En la sección 2.1 se ven las formas de resumir datos cualitativos.

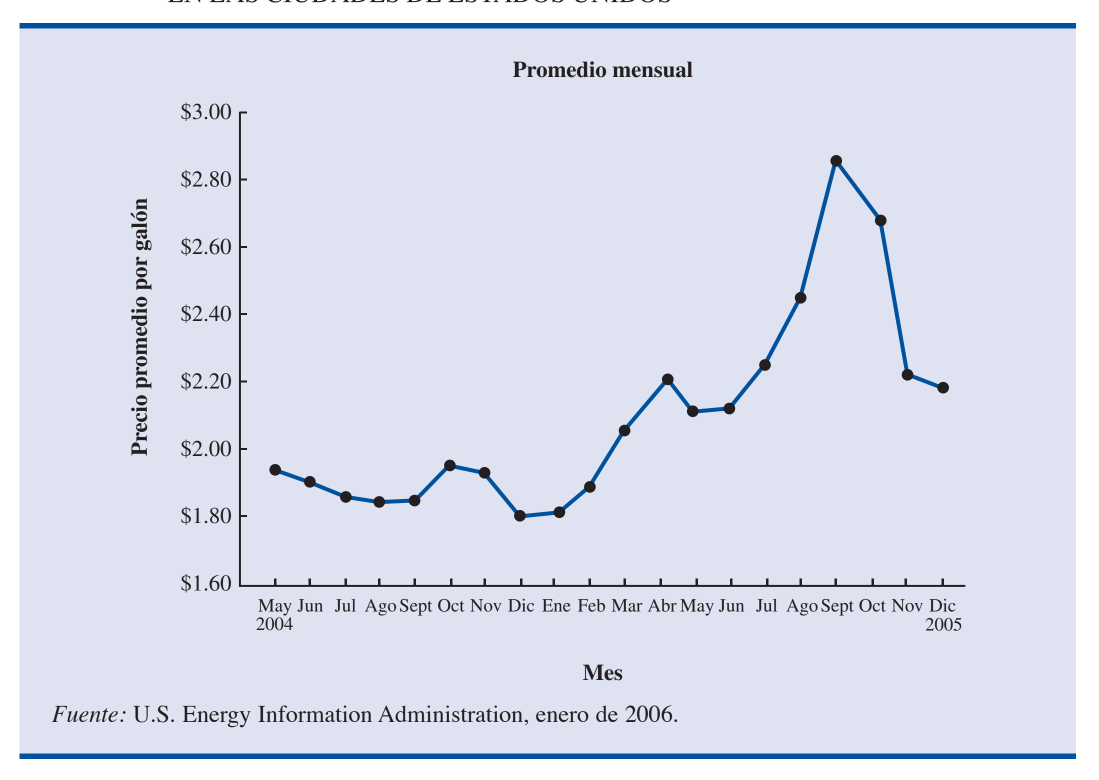
Por otro lado, las operaciones aritméticas sí tienen sentido en las variables cuantitativas. Por ejemplo, cuando se tienen variables cuantitativas, los datos se pueden sumar y luego dividir entre el número de observaciones para calcular el valor promedio. Este promedio suele ser útil y fácil de interpretar. En general hay más alternativas para el análisis estadístico cuando se tienen datos cuantitativos. La sección 2.2 y el capítulo 3 proporcionan condiciones para resumir datos cuantitativos.

Datos de sección transversal y de series de tiempo

Para los propósitos del análisis estadístico la distinción entre datos transversales y datos de series de tiempo es importante. **Datos de sección transversal** son los obtenidos en el mismo o aproximadamente el mismo momento (punto en el tiempo). Los datos de la tabla 1.1 son datos transversales porque describen las cinco variables de las 25 empresas del 25 S&P en un mismo momento. Los **datos de series de tiempo** son datos obtenidos a lo largo de varios periodos. Por ejemplo, la figura 1.1 presenta una gráfica de los precios promedio por galón de gasolina normal en las ciudades de Estados Unidos. En la gráfica se observa que los precios son bastantes estables entre $1.80 y $2.00 desde mayo de 2004 hasta febrero de 2005. Después el precio de la gasolina se vuelve volátil. Se eleva en forma notable culminando en un agudo pico en septiembre de 2005.

En las publicaciones sobre negocios y economía se encuentran con frecuencia gráficas de series de tiempo. Estas gráficas ayudan a los analistas a entender lo que ocurrió en el pasado, a identificar cualquier tendencia en el transcurso del tiempo y a proyectar niveles futuros para la series de tiempo. Las gráficas de datos de series de tiempo toman formas diversas como se muestra en la figura 1.2. Con un poco de estudio, estas gráficas suelen ser fáciles de entender y de interpretar.

FIGURA 1.1 PRECIO PROMEDIO POR GALÓN DE GASOLINA NORMAL EN LAS CIUDADES DE ESTADOS UNIDOS



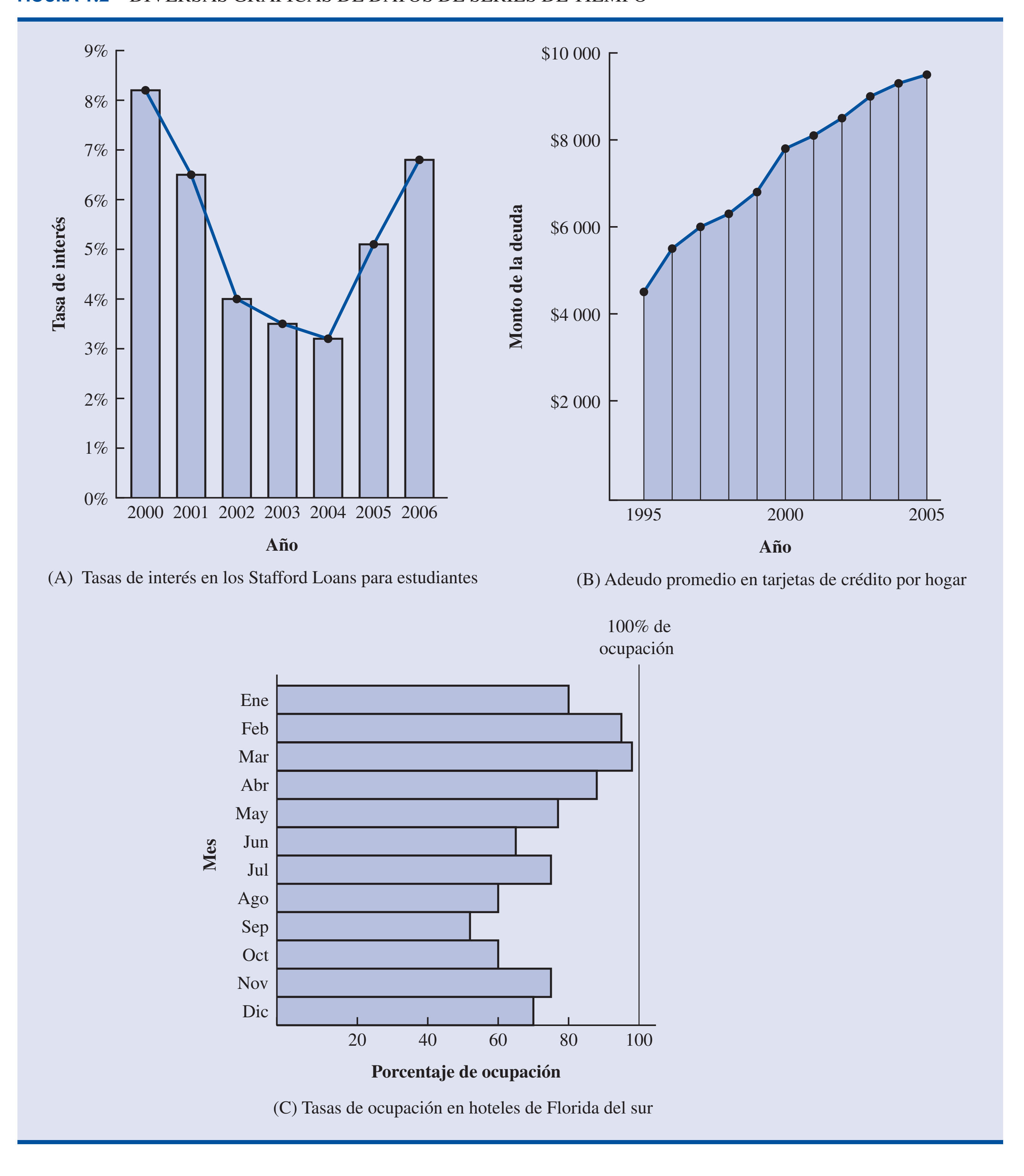
Fuente: U.S. Energy Information Administration, enero de 2006.

Por ejemplo, la gráfica (A) de la figura 1.2, muestra las tasas de interés en Stafford Loans para los estudiantes entre el año 2000 y el 2006. Después del año 2000 las tasas de interés disminuyen y llegan al nivel más bajo, 3.2%, en el año 2004. Pero, después de este año se observa un marcado aumento en estas tasas de interés, y llegan a 6.8% en el año 2006. El Departamento de Educación de Estados Unidos estima que más de 50% de los estudiantes terminan sus estudios con una deuda; esta creciente tasa de interés es una gran carga financiera para muchos estudiantes recién egresados.

En la gráfica (B) se observa un inquietante aumento en el adeudo promedio por hogar en tarjetas de crédito durante un periodo de 10 años, de 1995 a 2005. Advierta cómo en la series de tiempo se nota un aumento anual casi constante en el adeudo promedio por hogar en tarjetas de crédito que va de $4500 en 1995 a $9500 en 2005. En 2005 un adeudo promedio de 10 000 no parece lejano. La mayor parte de las empresas de tarjetas de crédito ofrecen tasas de interés iniciales relativamente bajas. Sin embargo, después de este periodo inicial, tasas de interés anuales del 18%, 20% y más son frecuentes. Estas tasas dificultan a los hogares pagar los adeudos de las tarjetas de crédito.

En la gráfica (C) se observan las tasas de ocupación en los hoteles de Florida del sur durante un año. Observe que la forma de esta gráfica es diferente a (A) y (B); en esta gráfica el tiempo en meses se encuentra en el eje vertical y no en el horizontal. Las tasas de ocupación más altas, 95% y 98%, se encuentran en los meses de febrero y marzo que es cuando el clima en Florida del sur es atractivo para los turistas. En efecto, de enero a abril es la estación de mayor ocupación en los hoteles de Florida del sur. Por otro lado, las tasas de ocupación más bajas se observan de agosto a octubre, siendo la menor ocupación en septiembre. Las temperaturas demasiado elevadas y la estación de huracanes son las principales razones de la caída de la ocupación en este periodo.

FIGURA 1.2 DIVERSAS GRÁFICAS DE DATOS DE SERIES DE TIEMPO



(A) Tasas de interés en los Stafford Loans para estudiantes

(B) Adeudo promedio en tarjetas de crédito por hogar

(C) Tasas de ocupación en hoteles de Florida del sur

Las series de tiempo y los pronósticos con series de tiempo se verán en el capítulo 16 cuando se estudien los métodos de pronóstico. Fuera del capítulo 16, los métodos estadísticos que se presentan en este libro son para datos de sección transversal y no para series de tiempo

NOTAS Y COMENTARIOS

1. Una observación es el conjunto de mediciones obtenidas para cada elemento de un conjunto de datos. Por tanto, el número de observaciones es siempre igual al número de elementos. El número de mediciones de cada elemento es igual al número de variables. Entonces, el número total de datos se determina multiplicando el número de observaciones por el número de variables.
2. Los datos cuantitativos son discretos o continuos. Datos cuantitativos que miden cuántos (por ejemplo, el número de llamadas recibidas en 5 minutos) son discretos. Datos cuantitativos que miden cuánto (por ejemplo, peso o tiempo) son continuos porque entre los posibles valores de los datos no hay separación.



1.3 Fuentes de datos

Los datos se obtienen de fuentes ya existentes o por medio de encuestas y estudios experimentales realizados con objeto de recolectar nuevos datos.

Fuentes existentes

En algunos casos los datos que se necesitan para una determinada aplicación ya existen. Las empresas cuentan con diversas bases de datos sobre sus empleados, clientes y operaciones de negocios. Datos sobre los salarios de los empleados, sus edades y los años de experiencia suelen obtenerse de los registros internos del personal. Otros registros internos contienen datos sobre ventas, gastos de publicidad, costos de distribución, inventario y cantidades de producción. La mayor parte de las empresas cuentan también con datos detallados de sus clientes. En la tabla 1.2 se muestran algunos de los datos obtenibles de los registros internos de las empresas.

De las organizaciones que se especializan en la recolección y almacenamiento de datos se obtienen cantidades importantes de datos económicos y de negocios. Las empresas disponen de estas fuentes externas de datos si los compran o mediante acuerdos de arrendamiento con opción de compra. Tres empresas que proporcionan amplios servicios de bases de datos a clientes son Dun & Bradstreet, Bloomberg y Dow Jones & Company. ACNielsen e Information Resources, Inc. han hecho un exitoso negocio recolectando y procesando datos que venden a publicistas y a fabricantes de productos.

TABLA 1.2 EJEMPLOS DE DATOS DISPONIBLES DE LOS REGISTROS DE EMPRESAS INTERNACIONALES

Fuente	Algunos de los datos disponibles
Registros sobre los empleados	Nombre, dirección, número de seguridad social, salario, días de vacaciones, días de enfermedad y bonos
Registros de producción	Parte o número de producto, cantidad producida, costo de mano de obra y costo de materiales
Registros de inventario	Parte o número de producto, cantidad de unidades disponibles, nivel de reaprovisionamiento, cantidad económica a ordenar y programa de descuento
Registros de ventas	Número del producto, volumen de ventas, volumen de ventas por región y volumen de ventas por tipo de cliente
Registros de créditos	Nombre del cliente, dirección, número de teléfono, crédito límite y cuentas por cobrar
Perfil de clientes	Edad, género, nivel de ingresos, número de miembros en la familia, dirección y preferencias

También se obtienen datos de diversas asociaciones industriales y de organizaciones de interés especial. La asociación Travel Industry Association of America cuenta con información relacionada con los viajes como número de turistas y gastos en viajes por estado. Estos datos interesan a empresas e individuos de la industria turística. El Graduate Management Admission Council cuenta con datos sobre calificaciones en exámenes, características de los estudiantes y programas de educación para administradores/directivos. La mayor parte de los datos de estas fuentes están a disposición de los usuarios calificados a un costo moderado.

La importancia de Internet como fuente de datos y de información estadística sigue creciendo. Casi todas las empresas cuentan con una página Web que proporciona información general acerca de la empresa así como datos sobre ventas, cantidad de empleados, cantidad de productos, precios de los productos y especificaciones de los productos. Además, muchas empresas se especializan ahora en proporcionar información a través de Internet. Con lo que uno puede tener acceso a cotizaciones de acciones, precios de comidas en restaurantes, datos de salarios y a una variedad casi infinita de información.

Las dependencias de los gobiernos son otra fuente importante de datos. Por ejemplo, el Departamento del Trabajo de Estados Unidos cuenta con una cantidad considerable de datos sobre tasas de empleo, tasas de salarios, magnitud de la fuerza laboral y pertenencia a sindicatos. En la tabla 1.3 se presentan algunas de las dependencias de gobierno junto con los datos que proporcionan. La mayor parte de las dependencias de los gobiernos que recolectan y procesan datos también los ponen a disposición a través de una página en la Web. Por ejemplo, la Oficina de Censos de Estados Unidos tiene una abundancia de datos en el sitio www.census.gov. En la figura 1.3 se muestra la página Web de la Oficina de Censos de Estados Unidos.

Estudios estadísticos

Algunas veces, los datos necesarios para una aplicación particular no se pueden obtener de las fuentes existentes. En tales casos los datos suelen conseguirse realizando un estudio estadístico. Dichos estudios se clasifican como *experimentales* u *observacionales*.

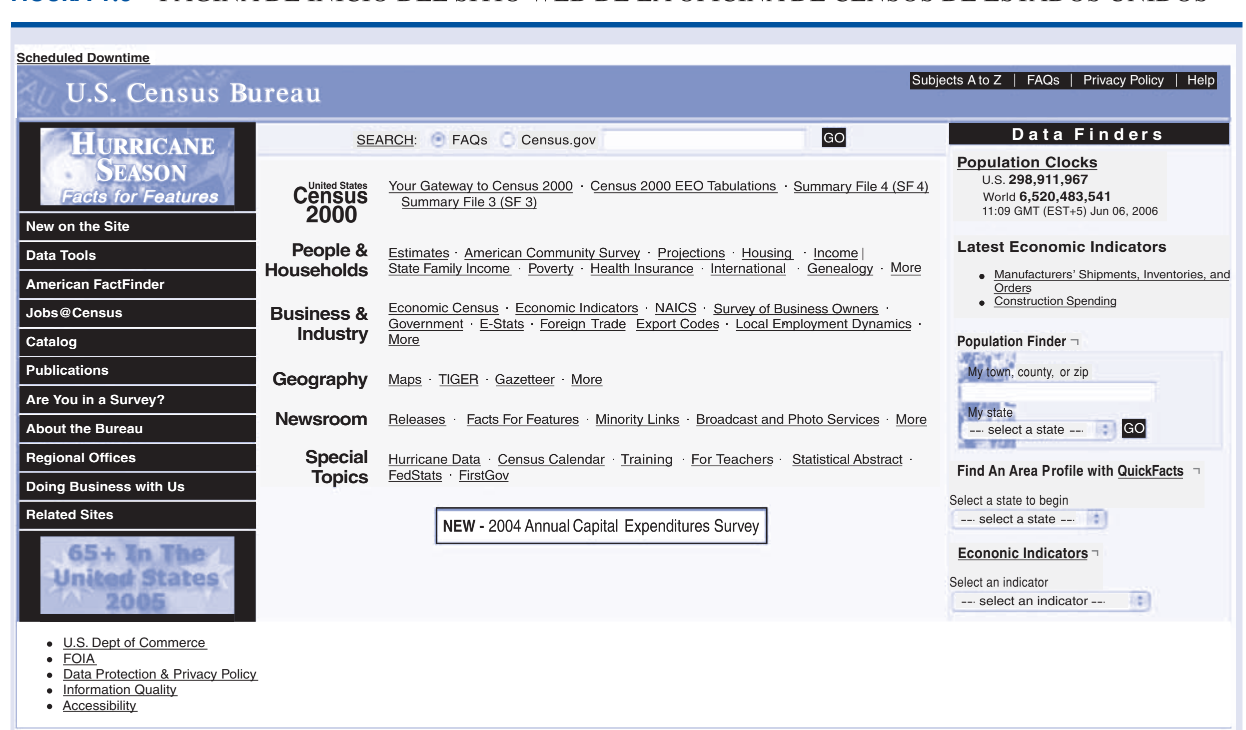
El mayor estudio estadístico experimental jamás realizado se cree que es el experimento del Servicio de Salud Pública para la vacuna Salk contra la polio. Se eligieron casi 2 millones de niños de 1o., 2o. y 3er. grados en Estados Unidos.

En los estudios experimentales se identifica primero la variable de interés. Después se ubica otra u otras variables que son controladas para lograr datos de cómo ésta influye sobre la variable de interés. Por ejemplo, a una empresa farmacéutica le interesa realizar un experimento para saber la forma en que un medicamento afecta la presión sanguínea. La variable que interesa en el estudio es la presión sanguínea. Otra variable es la dosis del nuevo medicamento que se espera tenga un efecto causal sobre la presión sanguínea. Para obtener estos datos acerca del nuevo medicamento, los investigadores eligen una muestra de individuos. La dosis del medicamento se controla dando diferentes dosis a distintos grupos de individuos. Antes y después se mide la pre-

TABLA 1.3 EJEMPLO DE LOS DATOS DISPONIBLES DE ALGUNAS DEPENDENCIAS GUBERNAMENTALES

Dependencia gubernamental	Algunos de los datos disponibles
Oficina de Censos *www.census.gov*	Datos poblacionales, número de hogares e ingresos de los hogares
Junta de la Reserva Federal *www.federalreserve.gov*	Datos sobre dinero en circulación, créditos a plazos, tasas de cambio y tasas de interés
Oficina de Administración y Presupuesto *www.whitehouse.gov/omb*	Datos sobre ingresos, gastos y deudas del gobierno federal
Departamento de Comercio *www.doc.gov*	Datos sobre las actividades comerciales, valor de los embarques por industria, nivel de ganancia por industria e industrias en crecimiento y en decremento
Oficina de Estadística Laboral *www.bls.gov*	Gasto de los consumidores, salarios por hora, tasa de desempleo y estadísticas internacionales

FIGURA 1.3 PÁGINA DE INICIO DEL SITIO WEB DE LA OFICINA DE CENSOS DE ESTADOS UNIDOS



sión sanguínea en cada grupo. El análisis estadístico de los datos experimentales ayuda a determinar el efecto del nuevo medicamento sobre la presión sanguínea.

Los estudios sobre fumadores y no fumadores son estudios observacionales porque los investigadores no determinan o controlan quién fuma y quién no.

En los estudios estadísticos no experimentales y observacionales, no se controlan las variables de interés. El tipo más usual de estudio observacional es quizá una encuesta. Por ejemplo, en una encuesta mediante entrevistas personales, primero se identifican las preguntas de la investigación. Después se presenta un cuestionario a los individuos de la muestra. Algunos restaurantes emplean estudios observacionales para obtener datos acerca de la opinión de sus clientes respecto a la calidad de los alimentos, del servicio, de la atmósfera, etc. En la figura 1.4 se presenta un cuestionario empleado por el restaurante Lobster Pot de Florida. Observe que en el cuestionario se pide a los clientes evaluar cinco variables: calidad de los alimentos, amabilidad en el servicio, prontitud en el servicio, limpieza y gestión. Las categorías para las respuestas de excelente, bueno, satisfactorio e insatisfactorio proporcionan datos ordinales que permiten a los directivos de Lobster Pot evaluar la calidad de operación del restaurante.

Los directivos que deseen emplear datos y análisis estadístico como ayuda en la toma de decisiones deben estar conscientes del tiempo y costo que requiere la obtención de los datos. Cuando es necesario obtener los datos en poco tiempo, es deseable el uso de fuentes de datos ya existentes. Si no es posible obtener con facilidad datos importantes de fuentes ya existentes, debe tomarse en cuenta el tiempo y el costo necesarios para obtener los datos. En todos los casos, las personas encargadas de tomar las decisiones deben considerar la contribución del análisis estadístico en el proceso de la toma de decisiones. El costo de la adquisición de datos y del subsiguiente análisis no deben exceder a los ahorros generados por el uso de esta información para tomar una decisión mejor.

Errores en la adquisición de datos

Los directivos siempre deben estar conscientes de la posibilidad de errores en los datos de los estudios estadísticos. Usar datos erróneos es peor que no usar ningún dato. Un error en la adquisición de datos se tiene siempre que el valor del dato obtenido no es igual al verdadero valor o al valor real que se hubiera obtenido con un procedimiento correcto. Estos errores ocurren de va-

FIGURA 1.4 CUESTIONARIO PARA CONOCER LA OPINIÓN DE LOS CLIENTES EMPLEADO EN EL RESTAURANTE THE LOBSTER POT DE REDINGTON SHORES, FLORIDA

The LOBSTER Pot
RESTAURANT

*N*os alegramos de su visita al restaurante Lobster Pot y queremos estar seguros de que volverá. De manera que si tiene unos minutos le agradeceríamos mucho que nos llenara esta tarjeta. Sus comentarios y sugerencias son extremadamente importantes para nosotros. Gracias.

Nombre de la persona que lo atendió ________________

	Excelente	Bueno	Satisfactorio	Insatisfactorio
Calidad de los alimentos	❑	❑	❑	❑
Amabilidad en el servicio	❑	❑	❑	❑
Prontitud en el servicio	❑	❑	❑	❑
Limpieza	❑	❑	❑	❑
Gestión	❑	❑	❑	❑

Comentarios ____

____

¿Qué lo motivó a visitarnos? ____

____

Favor de depositarlo en el buzón de sugerencias que se encuentra a la entrada.

rias maneras. Por ejemplo, un entrevistador puede cometer un error de escritura, como una transposición al escribir la edad de una persona y en lugar de 24 años escribir 42 años, o en una entrevista, el entrevistado puede malinterpretar una pregunta y dar una respuesta incorrecta.

Los analistas de datos con experiencia tienen sumo cuidado tanto al recolectar los datos como al registrarlos para garantizar que no se cometan errores. Para comprobar la consistencia interna de los datos se emplean procedimientos especiales. Tales procedimientos indican al analista, por ejemplo, que debe revisar la consistencia de los datos cuando un entrevistado aparece con 22 años de edad pero informa tener 20 años de experiencia en el trabajo. El analista de datos también debe revisar datos que tengan valores inusualmente grande o pequeños, llamados observaciones atípicas, que son candidatos a posibles errores en los datos. En el capítulo 3 se muestran algunos de los métodos estadísticos útiles para identificar observaciones atípicas.

Los errores suelen presentarse durante la adquisición de datos. Emplear a ciegas cualquier dato que se tenga o valerse de datos que fueron adquiridos con poco cuidado da como resultado información desorientadora y malas decisiones. Así, tomar medidas para adquirir datos precisos ayuda a garantizar información confiable y valiosa para la toma de decisiones.

1.4 Estadística descriptiva

La mayor parte de la información estadística en periódicos, revistas, informes de empresas y otras publicaciones consta de datos que se resumen y presentan en una forma fácil de leer y de entender. A estos resúmenes de datos, que pueden ser tabulares, gráficos o numéricos se les conoce como **estadística descriptiva**.

TABLA 1.4 FRECUENCIAS Y FRECUENCIAS PORCENTUALES DE LA VARIABLE BOLSA DE VALORES

Bolsa de valores	Frecuencia	Frecuencia porcentual
Bolsa de Nueva York	20	80
Mercado Nacional Nasdaq	5	20
Totales	25	100

Vuelva al conjunto de datos de la tabla 1.1 que presenta 25 de las empresas de S&P 500. Los métodos de la estadística descriptiva pueden emplearse para resumir la información en este conjunto de datos. Por ejemplo, en la tabla 1.4 se presenta un resumen tabular de los datos de la variable bolsa de valores. Un resumen gráfico de los mismos datos, al que se le llama gráfica de barras aparece en la figura 1.5. Estos tipos de resúmenes, tabular y gráfico, permiten que los datos sean más fáciles de interpretar. Al revisar la tabla 1.4 y la figura 1.5 es fácil entender que la mayor parte de las acciones del conjunto de datos cotizan en la bolsa de Nueva York. Si emplea porcentajes: 80% cotizan en la bolsa de Nueva York y 20% en el Nasdaq.

En la figura 1.6 se presenta un resumen gráfico, llamado histograma, de los datos de la variable cuantitativa precio por acción. El histograma facilita ver que los precios por acción van de $0 a $100, con una mayor concentración entre $20 y $60.

Además de las presentaciones tabular y gráfica para resumir datos se emplea también la estadística descriptiva numérica. El estadístico descriptivo más común para resumir datos es el promedio o media. Mediante los datos de la variable ganancia por acción de las acciones S&P de la tabla 1.1, el promedio se calcula sumando las ganancias por acción de las 25 acciones y dividien-

FIGURA 1.5 GRÁFICA DE BARRAS DE LA VARIABLE BOLSA DE VALORES

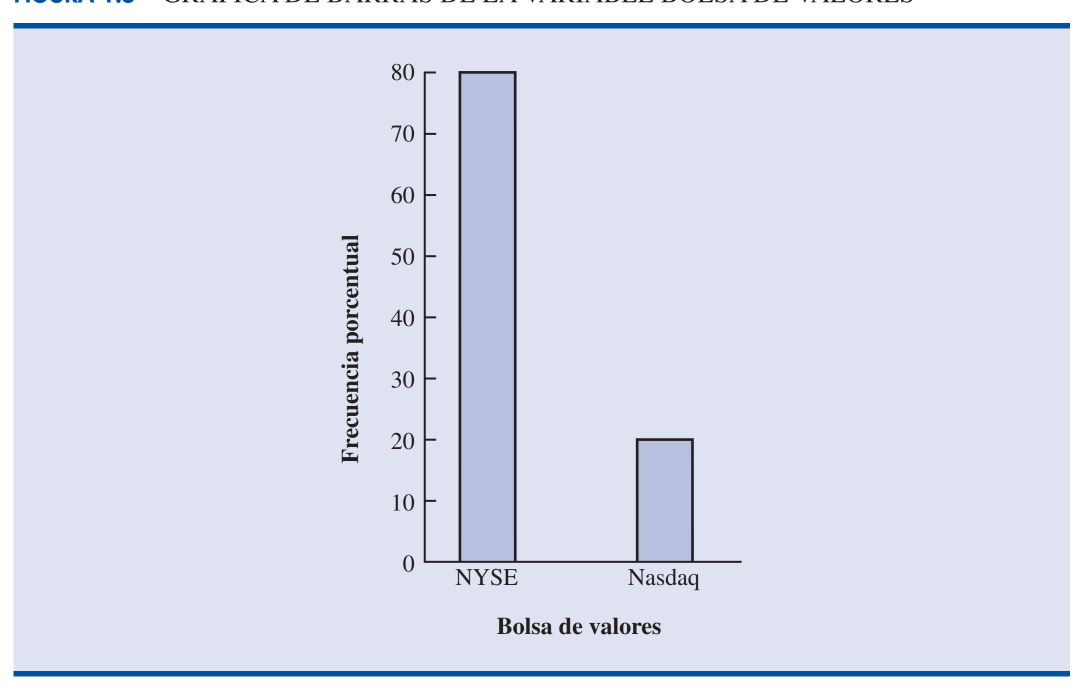
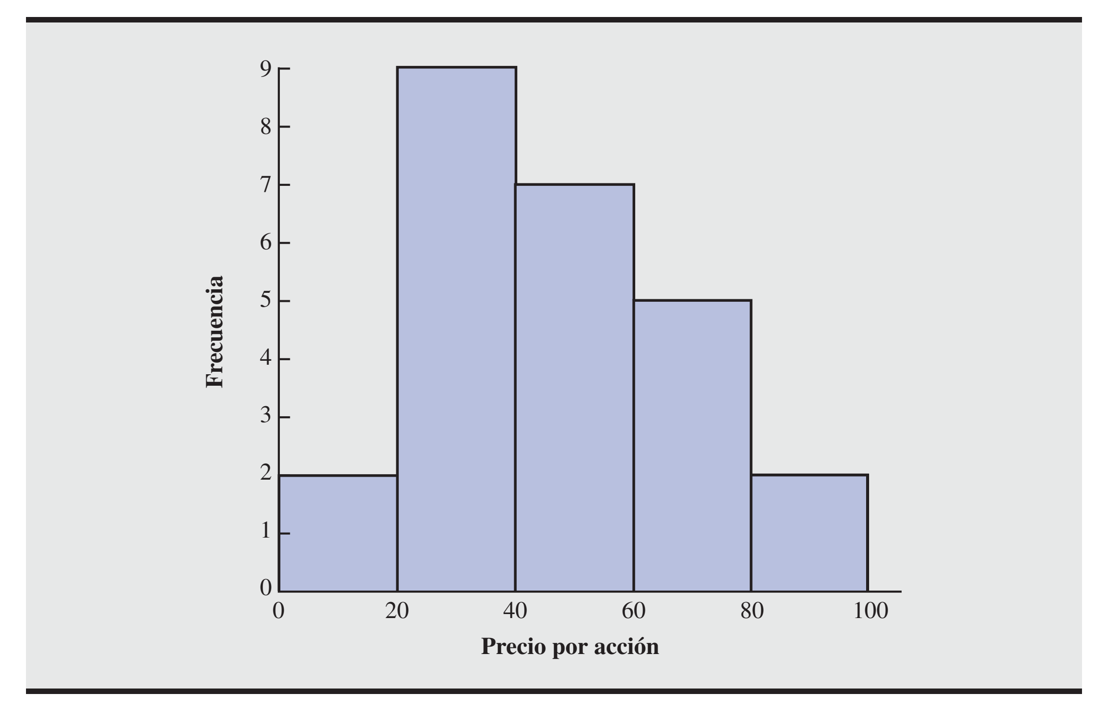


FIGURA 1.6 HISTOGRAMA DE LOS PRECIOS POR ACCIÓN DE 25 ACCIONES S&P



do entre 25. Al hacer esto se obtiene como ganancia promedio por acción $2.49. Este promedio da una tendencia central, o posición central, de los datos de la variable.

En numerosos campos sigue creciendo el interés por los métodos estadísticos que son aplicables para elaborar y presentar estadísticas descriptivas. En los capítulos 2 y 3 se dedica la atención a los métodos tabulares, gráficos y numéricos de la estadística descriptiva.



1.5 Inferencia estadística

En muchas situaciones se requiere información acerca de grupos grandes de elementos (individuos, empresas, votantes, hogares, productos, clientes, etc.). Pero, debido al tiempo, costo y a otras consideraciones, sólo es posible recolectar los datos de una pequeña parte de este grupo. Al grupo grande de elementos en un determinado estudio se le llama **población** y al grupo pequeño **muestra**. En términos formales se emplean las definiciones siguientes.

POBLACIÓN

La población es el conjunto de todos los elementos de interés en un estudio determinado.

MUESTRA

La muestra es un subconjunto de la población.

El gobierno de Estados Unidos realiza un censo cada 10 años. Las empresas de investigación de mercado realizan estudios muestrales cada día.

Al proceso de realizar un estudio para recolectar datos de toda una población se le llama **censo**. Al proceso de efectuar un estudio para recolectar datos de una muestra se le llama **encuesta muestral**. Una de las principales contribuciones de la estadística es emplear datos de una muestra para hacer estimaciones y probar hipótesis acerca de las características de una población mediante un proceso al que se le conoce como **inferencia estadística**.

Como un ejemplo de inferencia estadística, considere un estudio realizado por Norris Electronics. Norris fabrica focos de alta intensidad que se emplean en diversos productos electrónicos. Con objeto de incrementar la vida útil de estos focos, el grupo de diseño del producto elaboró un filamento nuevo. En este caso, la población está definida por todos los focos que se produzcan con el filamento nuevo. Para evaluar las ventajas del filamento, se fabricaron 200 focos. Los datos recolectados de esta muestra dan el número de horas que duró cada foco hasta que se quemara el filamento. Véase la tabla 1.5.

Suponga que Norris desea usar estos datos muestrales para hacer una inferencia acerca del número de horas promedio de vida útil de todos los focos que se producen con el filamento nuevo. Al sumar los 200 valores de la tabla 1.5 y dividir la suma entre 200 se obtiene el promedio del tiempo de vida de los focos: 76 horas. Este resultado muestral sirve para estimar que el tiempo de vida promedio de los focos de la población es 76 horas. En la figura 1.7 se proporciona un resumen gráfico del proceso de inferencia estadística empleado por Norris Electronics.

Siempre que un estadístico usa una muestra para estimar una característica poblacional que interesa, suele proporcionar información acerca de la calidad o precisión de la estimación. En el ejemplo de Norris, el estadístico puede informar que la estimación puntual del tiempo de vida promedio de la población de los nuevos focos es 76 horas con un margen de error de ±4 horas. Entonces, el intervalo de estimación del tiempo de vida promedio de los focos fabricados con el nuevo filamento es de 72 a 80 horas. El estadístico también puede informar qué tan confiado está de que el intervalo de 72 a 80 horas contenga el promedio poblacional.



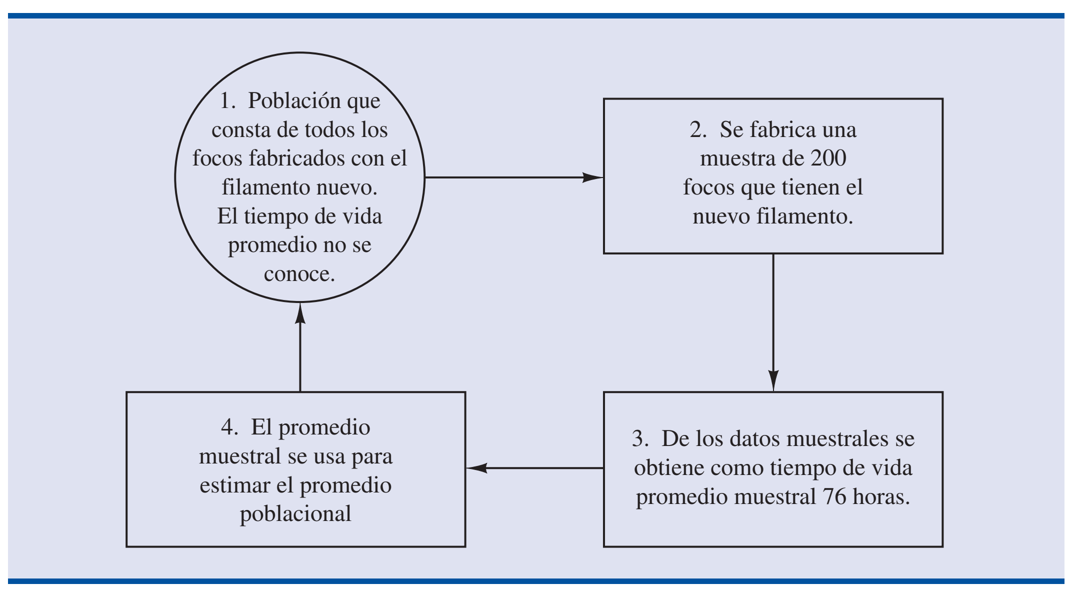
Norris

TABLA 1.5 HORAS DE DURACIÓN DE UNA MUESTRA DE 200 FOCOS DE NORRIS

107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81
75	90	62	89	71	71	74	70	74	70
65	81	75	62	94	71	85	84	83	63
81	62	79	83	93	61	65	62	92	65
83	70	70	81	77	72	84	67	59	58
78	66	66	94	77	63	66	75	68	76
90	78	71	101	78	43	59	67	61	71
96	75	64	76	72	77	74	65	82	86
66	86	96	89	81	71	85	99	59	92
68	72	77	60	87	84	75	77	51	45
85	67	87	80	84	93	69	76	89	75
83	68	72	67	92	89	82	96	77	102
74	91	76	83	66	68	61	73	72	76
73	77	79	94	63	59	62	71	81	65
73	63	63	89	82	64	85	92	64	73



FIGURA 1.7 PROCESO DE INFERENCIA ESTADÍSTICA EMPLEADO EN EL EJEMPLO DE NORRIS ELECTRONICS



1.6 Las computadoras y el análisis estadístico

Como en el análisis estadístico suelen emplearse grandes cantidades de datos, los analistas usan software para realizar estos trabajos. Por ejemplo, calcular el tiempo de vida promedio de los 200 focos del ejemplo de Norris Electronics (véase tabla 1.5) resultaría muy tedioso si no se contara con una computadora. Para facilitar el uso de una computadora, los conjuntos de datos de este libro se proporcionan en el disco compacto que viene con el libro. Un logotipo al margen izquierdo del texto identifica a estos conjuntos de datos. Los archivos de datos se encuentran en formatos para Minitab y para Excel. Además, en los apéndices de los capítulos aparecen las instrucciones para llevar a cabo los procedimientos estadísticos usando Minitab y Excel.

Resumen

La estadística es el arte y la ciencia de recolectar, analizar, presentar e interpretar datos. Casi todos los estudiantes de áreas relacionadas con los negocios o la economía necesitan tomar un curso de estadística. Este libro empezó describiendo las aplicaciones típicas de la estadística a los negocios y a la economía.

Los datos consisten en hechos/informaciones y cifras que se recolectan y analizan. Las cuatro escalas de medición que se usan para obtener datos sobre una determinada variable son nominal, ordinal, de intervalo y de razón. La escala de medición para una variable es nominal cuando los datos son etiquetas o nombres que se usan para identificar un atributo de un elemento. La escala es ordinal si los datos presentan las propiedades de los datos nominales y tiene sentido hablar del orden o jerarquía de los datos. La escala es de intervalo si los datos presentan las propiedades de los datos ordinales y los intervalos entre valores se expresan en términos de una unidad fija de medición. Por último, la escala de medición es de razón si los datos presentan las propiedades de los datos de intervalo y tiene sentido hablar de la razón entre dos valores.

Para los propósitos del análisis estadístico, los datos son clasificables en cuantitativos y cualitativos. Los datos cualitativos emplean etiquetas o nombres para identificar un atributo en cada elemento. Los datos cualitativos emplean las escalas de medición nominal u ordinal y pueden ser no numéricos o numéricos. Los datos cuantitativos son valores numéricos que indican cuánto o cuántos. Los datos cuantitativos emplean las escalas de medición de intervalo o de razón. Las operaciones aritméticas usuales sólo tienen sentido si los datos son cuantitativos. Por tanto, los cálculos estadísticos usados para datos cuantitativos no siempre son apropiados para datos cualitativos.

En las secciones 1.4 y 1.5 se introdujeron los temas de estadística descriptiva e inferencia estadística. Estadística descriptiva son los métodos tabulares, gráficos o numéricos que se usan para resumir datos. El proceso de la inferencia estadística emplea los datos obtenidos de una muestra para hacer estimaciones o probar hipótesis acerca de las características de la población. En la última sección del capítulo se indicó que las computadoras facilitan el análisis estadístico. Los conjuntos de datos grandes en los archivos de Minitab o de Excel se encuentran en el disco compacto que va con el libro.

Glosario

Estadística El arte y la ciencia de recolectar, analizar, presentar e interpretar datos.
Datos Los hechos y las cifras que se recolectan, analizan y resumen para su presentación e interpretación.
Conjunto de datos Todos los datos recolectados en un estudio determinado.
Elementos Entidades sobre las que se recolectan los datos.
Variable Una característica que interesa de un elemento.
Observación El conjunto de mediciones obtenidas de un elemento determinado.
Escala nominal Escala de medición de una variable cuando los datos son etiquetas o nombres que se emplean para identificar un atributo de un elemento. Los datos nominales pueden ser no numéricos o numéricos.
Escala ordinal Escala de medición de una variable cuando los datos presentan las propiedades de los datos nominales y el orden o jerarquía de los datos tiene sentido. Los datos ordinales pueden ser no numéricos o numéricos.
Escala de intervalo Escala de medición de una variable cuando los datos presentan las propiedades de los datos ordinales y los intervalos entre valores se expresan en términos de una unidad o medida fija. Los datos de intervalo siempre son numéricos.
Escala de razón Escala de medición de una variable cuando los datos presentan todas las propiedades de los datos de intervalo y la razón entre dos valores tiene sentido. Los datos de razón siempre son numéricos.
Datos cualitativos Etiquetas o nombres utilizados para identificar un atributo de cada elemento. Los datos cualitativos usan las escalas de medición nominal y ordinal y pueden ser no numéricos o numéricos.
Datos cuantitativos Valores numéricos que indican cuánto o cuántos de algo. Los datos cuantitativos se obtienen mediante la escala de intervalo o de razón.
Variable cualitativa Una variable con datos cualitativos.
Variable cuantitativa Una variable con datos cuantitativos.
Datos de sección transversal Datos recolectados en el mismo o aproximadamente en el mismo momento.
Datos de series de tiempo Datos recolectados a lo largo de varios periodos de tiempo.
Estadística descriptiva Resúmenes tabulares, gráficos o numéricos de datos.
Población Conjunto de todos los elementos que interesan en un estudio determinado.
Muestra Un subconjunto de la población.
Censo Un estudio para recolectar los datos de toda la población.
Encuesta muestral Un estudio para recolectar los datos de una muestra.
Inferencia estadística El proceso de emplear los datos obtenidos de una muestra para hacer estimaciones o probar hipótesis acerca de las características de la población.

Ejercicios complementarios

1. Describa la diferencia entre estadística como dato numérico y estadística como disciplina o campo de estudio.

Autoexamen

2. La revista *Condé Nast Traveler* realiza una encuesta anual entre sus suscriptores con objeto de determinar los mejores alojamientos del mundo. En la tabla 1.6 se presenta una muestra de nueve hoteles europeos (*Condé Nast Traveler,* enero de 2000) Los precios de una habitación doble estándar van de $(precio más bajo) a $$$$(precio más alto). La calificación general corresponde a la evaluación de habitaciones, servicio, restaurante, ubicación/atmósfera y áreas públicas; cuanto más alta sea la calificación general, mayor es el nivel de satisfacción.
   a. ¿Cuántos elementos hay en este conjunto de datos?
   b. ¿Cuántas variables hay en este conjunto de datos?
   c. ¿Cuáles variables son cualitativas y cuáles cuantitativas?
   d. ¿Qué tipo de escala de medición se usa para cada variable?

Autoexamen

3. Vaya a la tabla 1.6.
   a. ¿Cuál es el número promedio de habitaciones en los nueve hoteles?
   b. Calcule la calificación general promedio.
   c. ¿Qué porcentaje de los hoteles se encuentra en Inglaterra?
   d. ¿En qué porcentaje de los hoteles el precio de la habitación es de $$?

4. Los equipos de sonido todo en uno, llamados minicomponentes, cuentan con sintonizador AM/FM, casetera doble, cargador para un disco compacto con bocinas separadas. En la tabla 1.7 se muestran los precios de menudeo, calidad de sonido, capacidad para discos compactos, sensibilidad y selectividad de la sintonización y cantidad de caseteras en los artículos de una muestra de 10 minicomponentes (*Consumer Report Buying Guide 2002*).
   a. ¿Cuántos elementos contiene este conjunto de datos?
   b. ¿Cuál es la población?
   c. Calcule el precio promedio en la muestra.
   d. Con los resultados del inciso c, estime el precio promedio para la población.

5. Considere el conjunto de datos de la muestra de los 10 minicomponentes que se muestra en la tabla 1.7.
   a. ¿Cuántas variables hay en este conjunto de datos?
   b. De estas variables, ¿cuáles son cualitativas y cuáles son cuantitativas?
   c. ¿Cuál es la capacidad promedio de CD en la muestra?
   d. ¿Qué porcentaje de los minicomponentes tienen una sintonización de FM buena o excelente?
   e. ¿Qué porcentaje de los minicomponentes tienen dos caseteras?

TABLA 1.6 CALIFICACIONES PARA NUEVE LUGARES DONDE ALOJARSE EN EUROPA



Hotel

Nombre del lugar	País	Precio de la habitación	Número de habitaciones	Calificación general
Graveteye Manor	Inglaterra	$$	18	83.6
Villa d'Este	Italia	$$$$	166	86.3
Hotel Prem	Alemania	$	54	77.8
Hotel d'Europe	Francia	$$	47	76.8
Palace Luzern	Suiza	$$	326	80.9
Royal Crescent Hotel	Inglaterra	$$$	45	73.7
Hotel Sacher	Austria	$$$	120	85.5
Duc de Bourgogne	Bélgica	$	10	76.9
Villa Gallici	Francia	$$	22	90.6

Fuente: Condé Nast Traveler, enero de 2000.

TABLA 1.7 UNA MUESTRA DE 10 MINICOMPONENTES



Marca y modelo	Precio ($)	Calidad de sonido	Capacidad para CD	Sintonización FM	Caseteras
Aiwa NSX-AJ800	250	Buena	3	Regular	2
JVC FS-SD1000	500	Buena	1	Muy buena	0
JVC MX-G50	200	Muy buena	3	Excelente	2
Panasonic SC-PM11	170	Regular	5	Muy buena	1
RCA RS 1283	170	Buena	3	Mala	0
Sharp CD-BA2600	150	Buena	3	Buena	2
Sony CHC-CL1	300	Muy buena	3	Muy buena	1
Sony MHC-NX1	500	Buena	5	Excelente	2
Yamaha GX-505	400	Muy buena	3	Excelente	1
Yamaha MCR-E100	500	Muy buena	1	Excelente	0

6. La Columbia House vende discos compactos a los miembros de su club de venta por correo. En una encuesta sobre música se les pidió a los nuevos miembros del club que llenaran un cuestionario con 11 preguntas. Algunas de las preguntas eran:
   a. ¿Cuántos discos compactos has comprado en los últimos 12 meses?
   b. ¿Eres miembro de algún club de venta de libros por correo (Sí o No)?
   c. ¿Cuál es tu edad?
   d. Incluyéndote a ti, de cuántas personas (adultos y niños) consta tu familia.
   e. ¿Qué tipo de música te interesa comprar? Se presentaban quince categorías entre las que se encontraban rock pesado, rock ligero, música contemporánea para adultos, rap y rancheras.

   Responde si los datos que se obtienen con cada pregunta son cualitativos o cuantitativos.

7. El hotel Ritz Carlton emplea un cuestionario de opinión del cliente para obtener datos sobre la calidad de sus servicios de restaurante y entretenimiento (The Ritz-Carlton Hotel, Naples, Florida, febrero de 2006). Se les pidió a los clientes que evaluaran seis puntos: recibimiento, servicio, alimentos, menú, atención y atmósfera. Los datos registrados para cada factor fueron 1 para Pasadero, 2 Regular, 3 Bueno y 4 Excelente.
   a. Las respuestas de los clientes proporcionan datos para seis variables. ¿Son estas variables cualitativas o cuantitativas?
   b. ¿Qué escala de medición se usa?

8. La empresa Gallup realizó una encuesta telefónica empleando una muestra aleatoria nacional compuesta de 1005 adultos de 18 años o más. En la encuesta se les preguntó a los participantes "Cómo considera que es su salud física en este momento" (www.gallup.com, 7 de febrero de 2002)". Las respuestas podían ser Excelente, Buena, Regular o Ninguna opinión.
   a. ¿Cuál es el tamaño de la muestra de esta investigación?
   b. ¿Son estos datos cualitativos o cuantitativos?
   c. ¿Sería conveniente usar promedios o porcentajes para resumir los datos de estas preguntas?
   d. De las personas que respondieron, 29% dijo que su salud era excelente. ¿Cuántos fueron los individuos que dieron esta respuesta?

9. El Departamento de Comercio informa haber recibido las siguientes solicitudes para concursar por el Malcolm Baldrige Nacional Quality Award: 23 de empresas fabricantes grandes, 18 de empresas grandes de servicios y 30 de negocios pequeños.
   a. ¿Es el tipo de empresa una variable cualitativa o cuantitativa?
   b. ¿Qué porcentaje de las solicitudes venían de negocios pequeños?

10. En una encuesta de *The Wall Street Journal* (13 de octubre de 2003) se les hacen a los suscriptores 46 preguntas acerca de sus características e intereses. De cada una de las preguntas si-

guientes, ¿cuál proporciona datos cualitativos o cuantitativos e indica la escala de medición apropiada?

a. ¿Cuál es su edad?
b. ¿Es usted hombre o mujer?
c. ¿Cuándo empezó a leer el *WSJ*? Preparatoria, universidad al comienzo de la carrera, a la mitad de la carrera, al final de la carrera o ya retirado.
d. ¿Cuánto tiempo hace que tiene su trabajo o cargo actual?
e. ¿Qué tipo de automóvil piensa comprarse la próxima vez que compre uno? Ocho categorías para las respuestas, entre las que se encontraban sedán, automóvil deportivo, miniván, etcétera.

11. Diga de cada una de las variables siguientes si es cualitativa o cuantitativa e indique la escala de medición a la que pertenece.
a. Ventas anuales.
b. Tamaño de los refrescos (pequeño, mediano, grande).
c. Clasificación como empleado (GS 1 a GS 18).
d. Ganancia por acción.
e. Modo de pago (al contado, cheque, tarjeta de crédito).

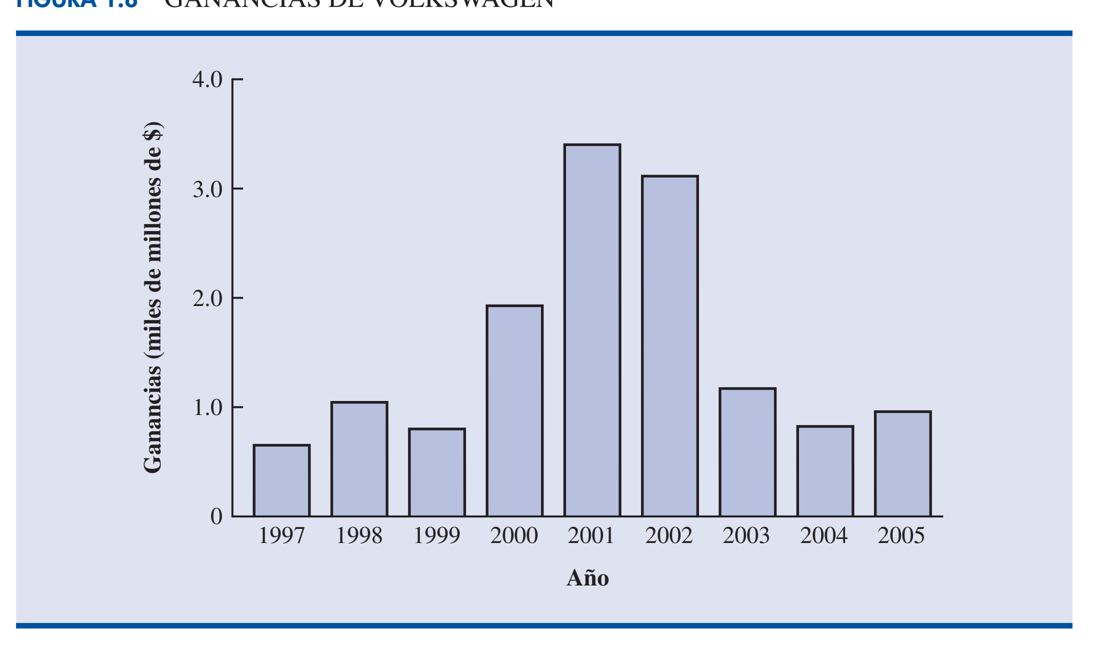
12. La Oficina de Visitantes a Hawai recolecta datos de los visitantes. Entre las 16 preguntas hechas a los pasajeros de un vuelo de llegada en junio de 2003 estaban las siguientes.
- Este viaje a Hawai es mi 1o., 2o., 3o., 4o. etc.
- La principal razón de este viaje es: (10 categorías para escoger entre las que se encontraban vacaciones, luna de miel, una convención).
- Dónde voy a alojarme: (11 categorías entre las que se encontraban hotel, departamento, parientes, acampar).
- Total de días en Hawai

a. ¿Cuál es la población que se estudia?
b. ¿El uso de un cuestionario es una buena manera de tener información de los pasajeros en los vuelos de llegada?
c. Diga de cada una de las cuatro preguntas si los datos que suministra son cualitativos o cuantitativos.



13. En la figura 1.8 se presenta una gráfica de barras que resume las ganancias de Volkswagen de los años 1997 a 2005 (*BusinessWeek,* 26 de diciembre de 2005).

FIGURA 1.8 GANANCIAS DE VOLKSWAGEN



a. ¿Estos son datos cualitativos o cuantitativos?
b. ¿Son datos de series de tiempo o datos de sección transversal?
c. ¿Cuál es la variable de interés?
d. Comente la tendencia en las ganancias de Volkswagen a lo largo del tiempo. El artículo de *BusinessWeek* (26 de diciembre de 2005) estimó las ganancias en 2006 en $600 millones o $0.6 mil millones. ¿Indica la figura si esta estimación parece ser razonable?
e. Un artículo similar que apareció en *BusinessWeek* el 23 de julio de 2001 sólo contaba con los datos de 1997 a 2000 junto con elevadas ganancias proyectadas para 2001. ¿Cómo era la perspectiva de las ganancias de Volkswagen en julio de 2001? En 2001, ¿parecía prometedor invertir en Volkswagen? Explique.
f. ¿Qué advertencia sugiere esta gráfica acerca de la proyección de datos como los de las ganancias de Volkswagen hacia el futuro?

14. CSM Worldwide pronostica la producción mundial de todos los fabricantes de automóviles. Los datos siguientes de CSM muestran el pronóstico de la producción mundial para General Motors, Ford, DaimlerChrysler y Toyota para los años 2004 a 2007 (*USA Today,* 21 de diciembre de 2005). Estos datos están dados en millones de vehículos.

Fabricante	**2004**	**2005**	**2006**	**2007**
General Motors	8.9	9.0	8.9	8.8
Ford	7.8	7.7	7.8	7.9
DaimlerChrysler	4.1	4.2	4.3	4.6
Toyota	7.8	8.3	9.1	9.6

a. Haga una gráfica de series de tiempo para los años 2004 a 2007 en la que se observe la cantidad de vehículos fabricados por cada empresa. Muestre las series de tiempo de los cuatro fabricantes en la misma gráfica.
b. General Motors ha sido sin discusión el principal fabricante de automóviles desde 1931. En esta gráfica de series de tiempo, ¿cuál es el mayor fabricante de automóviles? Explique.
c. Haga una gráfica que muestre los vehículos producidos por los fabricantes de automóviles usando los datos de 2007. ¿Está basada en datos de series de tiempo o en datos de sección transversal?

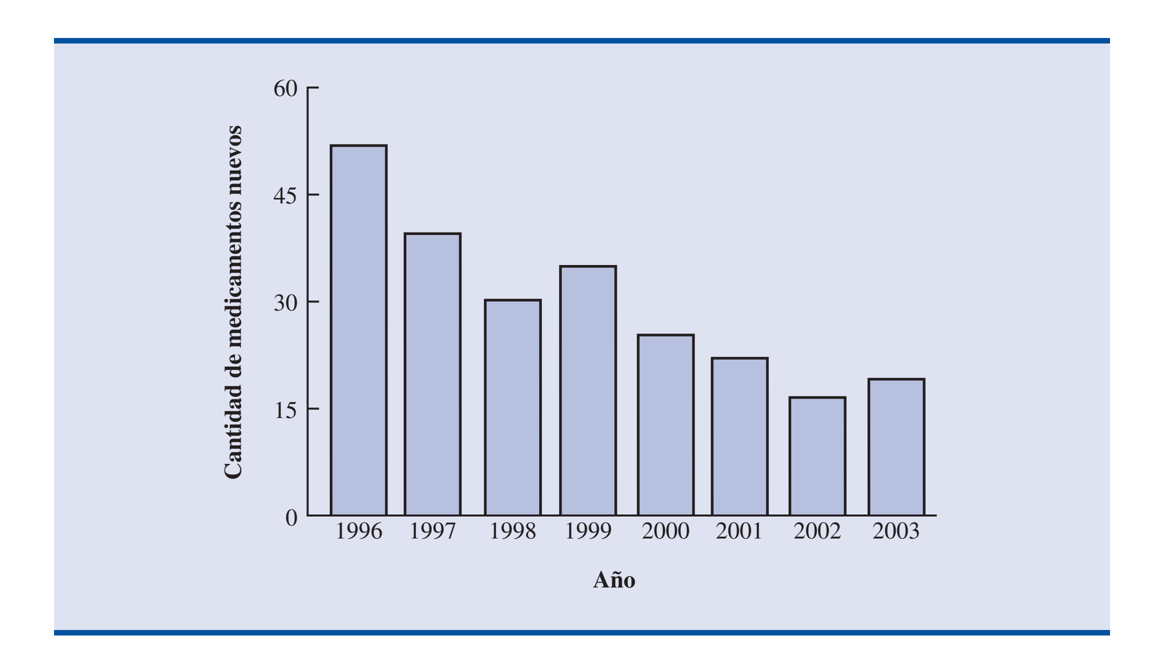
15. La Food and Drug Administration (FDA) da información sobre la cantidad de medicamentos aprobados en un periodo de ocho años (*The Wall Street Journal,* 12 de enero de 2004). En la figura 1.9 se presenta una gráfica de barras que resume el número de medicamentos nuevos aprobados cada año.
a. ¿Estos datos son cualitativos o cuantitativos?
b. ¿Son datos de series de tiempo o son datos de sección transversal?
c. ¿Cuántos medicamentos fueron aprobados en 2003?
d. ¿En qué año se aprobaron menos medicamentos? ¿Cuántos fueron?
e. Presente un comentario sobre la tendencia en el número de medicamentos nuevos aprobados por la FDA en este periodo de ocho años.

16. El departamento de marketing de su empresa elabora un refresco dietético que dice captará una gran parte del mercado de adultos jóvenes.
a. ¿Qué datos desearía ver antes de invertir una cantidad importante para introducir el nuevo producto en el mercado?
b. ¿Cómo esperaría que se obtuvieran los datos mencionados en el inciso a?

17. El directivo de una empresa grande recomienda un aumento de $10 000 para evitar que un empleado se cambie a otra empresa. ¿Qué fuentes de datos internas y externas pueden usarse para decidir si es apropiado ese incremento de salario?



FIGURA 1.9 NÚMERO DE MEDICAMENTOS NUEVOS APROBADOS POR LA FDA



18. En una encuesta a 430 viajeros de negocios se encontró que 155 de ellos empleaban los servicios de un agente de viajes para la preparación de sus viajes (*USA Today,* 20 de noviembre de 2003).
    a. Elabore una estadística descriptiva que sirva para estimar el porcentaje de viajeros de negocios que emplean un agente de viajes para preparar su viaje.
    b. Con la encuesta se encontró que la manera más frecuente en que los viajeros de negocios hacen los preparativos de su viaje es mediante un sitio en línea. Si 4% de los viajeros de negocios encuestados hacen los preparativos de su viaje de esta manera, ¿cuántos de los 430 encuestados emplearon un sitio en línea?
    c. Estos datos sobre cómo se hacen los preparativos, ¿son cualitativos o cuantitativos?
19. En un estudio sobre los suscriptores de *BusinessWeek* de Estados Unidos se recogen datos de una muestra de 2861 suscriptores. Cincuenta y nueve por ciento de los encuestados señalaron tener un ingreso de $75 000 o más y 50% indicaron poseer una tarjeta de crédito de American Express.
    a. ¿Cuál es la población de interés en este estudio?
    b. ¿Es el ingreso anual un dato cualitativo o cuantitativo?
    c. ¿Es la posesión de una tarjeta de crédito de American Express una variable cualitativa o cuantitativa?
    d. ¿Hacer este estudio requiere datos de series de tiempo o de sección transversal?
    e. Describa cualquier inferencia estadística posible para *BusinessWeek* con base en esta encuesta.
20. En una encuesta a 131 directores de inversión en Barron's se encontró lo siguiente (Barron's 28 de octubre de 2002):
    - De los dirigentes 43% se clasificaron como optimistas o muy optimistas sobre el mercado de acciones.
    - El rendimiento promedio esperado en los 12 meses siguientes en títulos de capital fue 11.2%.
    - La atención a la salud fue elegida por 21% como el sector con más probabilidad de ir a la cabeza del mercado en los próximos 12 meses.
    - Cuando se les preguntó cuánto tiempo se necesitaría para que las acciones de tecnología y telecomunicación recobraran un crecimiento sostenible, la respuesta promedio de los directivos fue 2.5 años.

a. Cite dos estadísticas descriptivas.
b. Haga una inferencia sobre la población de todos los directivos de inversiones respecto al rendimiento promedio esperado en los títulos de capital durante los siguientes 12 meses.
c. Haga una inferencia acerca de la cantidad de tiempo que se necesitará para que las acciones de tecnología y telecomunicación recobren un crecimiento sostenible.

21. En una investigación médica que duró siete años se encontró que las mujeres cuyas madres habían tomado el medicamento DES durante el embarazo, tenían el doble de posibilidades de presentar anormalidades en los tejidos que pudieran conducir a un cáncer, que aquellas cuyas madres no habían tomado este medicamento.
    a. En este estudio se compararon dos poblaciones. ¿Cuáles son?
    b. ¿Es posible pensar que los datos se obtuvieron mediante una encuesta o mediante un experimento?
    c. De la población de las mujeres cuyas madres habían tomado el medicamento DES durante el embarazo, se encontró que en una muestra de 3980 mujeres 63 presentaban anormalidades en tejidos que podrían conducir a un cáncer. Dé un estadístico descriptivo útil para estimar el número de mujeres, de cada 1000, de esta población que pueden presentan anormalidades en los tejidos.
    d. De la población de mujeres cuyas madres no tomaron el medicamento DES durante el embarazo, ¿cuál es el número estimado de mujeres, de cada 1000, que pueden presentar anormalidades en los tejidos?
    e. Estudios médicos a menudo utilizan muestras grandes (en este caso, 3980). ¿Por qué?

22. En otoño de 2003, Arnold Schwarzeneger disputó al gobernador Gray Davis la gobernatura de California. En una encuesta realizada entre los votantes registrados se encontró que Arnold Schwarzeneger iba a la cabeza con un porcentaje estimado de 54% (*Newsweek,* 8 de septiembre de 2003).
    a. ¿Cuál fue la población en este estudio?
    b. ¿Cuál fue la muestra en este estudio?
    c. ¿Por qué se empleó una muestra en esta situación? Explique.

23. Nielsen Media Research realiza cada semana un sondeo entre los televidentes de Estados Unidos y publica datos tanto de índice de audiencia como de participación en el mercado. El índice de audiencia de Nielsen es el porcentaje de hogares que tienen televisión y que están viendo un programa, mientras que la participación de Nielsen es el porcentaje de hogares que están viendo un programa, entre los hogares que tiene la televisión en uso. Por ejemplo, los resultados de Nielsen Media Research para la Serie Mundial de Béisbol de 2003 entre los Yankees de Nueva York y los Marlins de Florida dieron un índice de audiencia de 12.8% y una participación de 22% (Associated Press, 27 de octubre de 2003). Por tanto, 12.8% de los hogares que tenían televisión estaban viendo la Serie Mundial y 22% de los hogares que estaban viendo la televisión, estaban viendo la Serie Mundial. A partir de los datos de índices de audiencia y de participación, Nielsen publica un ranking semanal de los programas de televisión así como un ranking semanal de las cuatro principales cadenas de televisión en Estados Unidos: ABC, CBS, NBC y Fox.
    a. ¿Qué trata de medir Nielsen Media Research?
    b. ¿Cuál es la población?
    c. ¿Por qué se usaría una muestra en esta situación?
    d. ¿Qué tipo de decisiones o de acciones están basadas en los rankings de Nielsen?

24. En una muestra con cinco calificaciones de los estudiantes en un determinado examen los datos fueron: 72, 65, 82, 90, 76. ¿Cuáles de las afirmaciones siguientes son correctas y cuáles deben cuestionarse como una generalización excesiva?
    a. La calificación promedio de este examen en la muestra de las calificaciones de cinco estudiantes es 77.
    b. La calificación promedio de todos los estudiantes en este examen es 77.
    c. Una estimación para la calificación promedio de todos los estudiantes que hicieron el examen es 77.
    d. Más de la mitad de los estudiantes que hicieron en examen tendrán calificaciones entre 70 y 85.
    e. Si se incluyen en la muestra otros cinco estudiantes, sus calificaciones estarán entre 65 y 90.

TABLA 1.8 CONJUNTO DE DATOS DE 25 ACCIONES SHADOW

Empresa	Bolsa de valores	Denominación abreviada Symbol	Capacidad de mercado (millones de $)	Relación precio/ganancia	Margen de ganancia bruta (%)
DeWolfe Companies	AMEX	DWL	36.4	8.4	36.7
North Coast Energy	OTC	NCEB	52.5	6.2	59.3
Hansen Natural Corp.	OTC	HANS	41.1	14.6	44.8
MarineMax, Inc.	NYSE	HZO	111.5	7.2	23.8
Nanometrics Incorporated	OTC	NANO	228.6	38.0	53.3
TeamStaff, Inc.	OTC	TSTF	92.1	33.5	4.1
Environmental Tectonics	AMEX	ETC	51.1	35.8	35.9
Measurement Specialties	AMEX	MSS	101.8	26.8	37.6
SEMCO Energy, Inc.	NYSE	SEN	193.4	18.7	23.6
Party City Corporation	OTC	PCTY	97.2	15.9	36.4
Embrex, Inc.	OTC	EMBX	136.5	18.9	59.5
Tech/Ops Sevcon, Inc.	AMEX	TO	23.2	20.7	35.7
ARCADIS NV	OTC	ARCAF	173.4	8.8	9.6
Qiao Xing Universal Tele.	OTC	XING	64.3	22.1	30.8
Energy West Incorporated	OTC	EWST	29.1	9.7	16.3
Barnwell Industries, Inc.	AMEX	BRN	27.3	7.4	73.4
Innodata Corporation	OTC	INOD	66.1	11.0	29.6
Medical Action Industries	OTC	MDCI	137.1	26.9	30.6
Instrumentarium Corp.	OTC	INMRY	240.9	3.6	52.1
Petroleum Development	OTC	PETD	95.9	6.1	19.4
Drexler Technology Corp.	OTC	DRXR	233.6	45.6	53.6
Gerber Childrenswear Inc.	NYSE	GCW	126.9	7.9	25.8
Gaiam, Inc.	OTC	GAIA	295.5	68.2	60.7
Artesian Resources Corp.	OTC	ARTNA	62.8	20.5	45.5
York Water Company	OTC	YORW	92.2	22.9	74.2



Shadow02

25. En la tabla 1.8 aparece un conjunto de datos con información sobre 25 de las acciones shadow vigiladas por la American Association of Individual Investors (aaii.com, febrero de 2002). Acciones shadow son acciones comunes de empresas pequeñas que no son estrechamente vigiladas por los analistas de Wall Street. Este conjunto de datos se encuentra también en el disco compacto que se incluye en este libro, en el archivo Shadow02.
   a. ¿Cuántas variables hay en este conjunto de datos?
   b. ¿Qué variables son cualitativas y cuáles son cuantitativas?
   c. Par la variable bolsa de valores muestre la frecuencia y la frecuencia porcentual de AMEX, NYSE y OTC. Construya una gráfica de barras como la de la figura 1.5.
   d. Muestre la distribución de frecuencias del margen de ganancia bruta empleando cinco intervalos: 0–14.9, 15–29.9, 30–44.9. 45–59.9 y 60–74.9. Construya un histograma como el de la figura 1.6.
   e. ¿Cuál es la proporción precio/ganancia promedio?



CAPÍTULO 2

Estadística descriptiva: presentaciones tabulares y gráficas

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

LA EMPRESA COLGATE-PALMOLIVE*
NUEVA YORK, NUEVA YORK

La empresa Colgate-Palmolive empezó en la Ciudad de Nueva York en 1806 como una pequeña tienda de jabones y velas. Hoy, Colgate-Palmolive emplea más de 4000 personas que trabajan en 200 países y territorios del mundo. Aunque es más conocida por sus marcas Colgate, Palmolive, Ajax y Fab, la empresa comercializa los productos Mennen, Hill's Science Diet y Hill's Prescription Diet.

La empresa Colgate-Palmolive aplica la estadística en su programa de aseguramiento de la calidad en los detergentes caseros para la ropa. Le interesa la satisfacción del cliente con la cantidad de detergente en los paquetes. Todos los paquetes de cierto tamaño se llenan con la misma cantidad de detergente en peso, aunque el volumen del detergente varía de acuerdo con la densidad del polvo detergente. Por ejemplo, si la densidad del detergente es alta, se necesita una cantidad menor de detergente para tener el peso señalado en el paquete. El resultado es que cuando el cliente abre el paquete le parece que no ha sido bien llenado.

Para controlar el problema del peso del polvo de detergente, se han establecido límites en el nivel aceptable de la densidad del polvo. Con periodicidad se toman muestras estadísticas y se mide la densidad de la muestra de polvo. Los resúmenes de los datos se les proporcionan a los operarios para que de ser necesario lleven a cabo acciones correctivas, de manera que la densidad se mantenga dentro de las especificaciones de calidad establecidas.

En la tabla y figura adjuntas se presentan una distribución de frecuencia y un histograma obtenidos con 150 muestras tomadas en una semana. Densidades mayores a 0.40 son inaceptablemente altas. De acuerdo con la distribución de frecuencia y al histograma la operación satisface los lineamientos de calidad ya que todas las densidades son menores o iguales a 0.40. A la vista de estos resúmenes estadísticos los directivos estarán satisfechos con la calidad del proceso de producción de detergente.

En este capítulo se estudiarán métodos tabulares y gráficos de la estadística descriptiva como distribuciones de frecuencia, gráficas de barras, histogramas, diagramas de tallo y hoja, tabulaciones cruzadas y otros. El objeto de estos métodos es resumir los datos de manera que sean entendibles e interpretables con facilidad.

*Los autores agradecen a William R. Fawle, director de aseguramiento de la calidad de la empresa Colgate-Palmolive por proporcionarles este artículo para *La estadística en la práctica*.

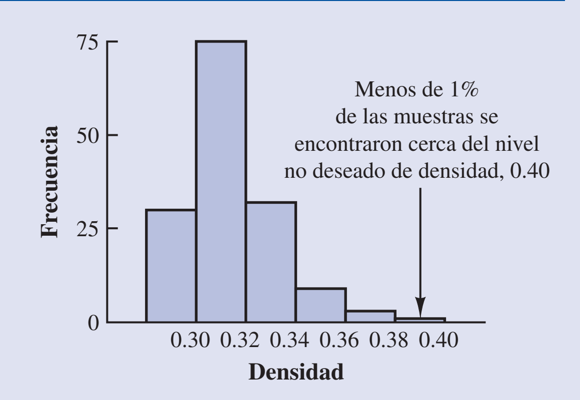


Los resúmenes estadísticos ayudan a mantener la calidad de estos productos de Colgate-Palmolive © Joe Higgins/South Western.

Distribución de frecuencia de los datos de densidad

Densidad	Frecuencia
0.29–0.30	30
0.31–0.32	75
0.33–0.34	32
0.35–0.36	9
0.37–0.38	3
0.39–0.40	1
Total	150

Histograma de los datos de densidad



Como se indicó en el capítulo 1, los datos se clasifican en cualitativos o cuantitativos. Los **datos cualitativos** emplean etiquetas o nombres para determinar categorías de elementos iguales. Los **datos cuantitativos** son números que indican cuánto o cuántos.

En este capítulo se presentan los métodos tabulares y gráficos empleados para datos cualitativos y cuantitativos. Los resúmenes gráficos o tabulares de datos se encuentran en reportes anuales, en artículos en los periódicos y en estudios de investigación. Todo mundo se encuentra con este tipo de presentaciones. Por tanto, es útil saber cómo se hacen y se interpretan. Se empezará con los métodos tabulares y gráficos para resumir datos que se refieren a una sola variable. En la última sección se introducen los métodos para resumir datos cuando lo que interesa es la relación entre dos variables.

Los paquetes modernos de software para estadística proporcionan muchas posibilidades para resumir datos y elaborar presentaciones gráficas. Minitab y Excel son dos paquetes muy empleados. En los apéndices de este capítulo se muestran algunas de sus posibilidades.



Resumen de datos cualitativos

Distribución de frecuencia

Conviene iniciar el estudio acerca del uso de los métodos tabulares y gráficos para resumir datos cualitativos con la definición de **distribución de frecuencia**.

DISTRIBUCIÓN DE FRECUENCIA

Una distribución de frecuencia es un resumen tabular de datos que muestra el número (frecuencia) de elementos en cada una de las diferentes clases disyuntas (que no se sobreponen).

Con el ejemplo siguiente se muestra la elaboración e interpretación de una distribución de frecuencia de datos cualitativos. Cinco refrescos muy conocidos son Coca cola clásica (Coke Classic), Coca cola de dieta (Diet Coke), Dr. Pepper, Pepsi y Sprite. Suponga que los datos de la tabla 2.1 muestran los refrescos que fueron comprados en una muestra de 50 ventas de refresco.

TABLA 2.1 DATOS DE UNA MUESTRA DE 50 VENTAS DE REFRESCO



Coke Classic	Sprite	Pepsi
Diet Coke	Coke Classic	Coke Classic
Pepsi	Diet Coke	Coke Classic
Diet Coke	Coke Classic	Coke Classic
Coke Classic	Diet Coke	Pepsi
Coke Classic	Coke Classic	Dr. Pepper
Dr. Pepper	Sprite	Coke Classic
Diet Coke	Pepsi	Diet Coke
Pepsi	Coke Classic	Pepsi
Pepsi	Coke Classic	Pepsi
Coke Classic	Coke Classic	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coke Classic	Coke Classic
Coke Classic	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coke Classic	Pepsi	Sprite
Coke Classic	Diet Coke	

TABLA 2.2

DISTRIBUCIÓN DE FRECUENCIA DE LAS VENTAS DE REFRESCO

Refresco	Frecuencia
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

Para elaborar una distribución de frecuencia con estos datos, se cuenta el número de veces que aparece cada refresco en la tabla 2.1. La Coca cola clásica (Coke Classic) aparece 19 veces, la Coca cola de dieta (Diet Coke) 8 veces, Dr. Pepper 5 veces, Pepsi 13 veces y Sprite 5 veces. Esto queda resumido en la distribución de frecuencia de la tabla 2.2.

Esta distribución de frecuencia proporciona un resumen de cómo se distribuyeron las 50 ventas entre los cinco refrescos. El resumen aporta más claridad que los datos originales de la tabla 2.1. Al observar esta distribución de frecuencia, es claro que Coca cola clásica es el refresco que más se vende, Pepsi el segundo, Coca cola de dieta el tercero y Sprite y Dr. Pepper están empatados en el cuarto lugar. La distribución de frecuencia resume la información sobre la popularidad de los cinco refrescos.

Distribuciones de frecuencia relativa y de frecuencia porcentual

En una distribución de frecuencia se aprecia el número (frecuencia) de los elementos de cada una de las diversas clases disyuntas. Sin embargo, con frecuencia lo que interesa es la proporción o porcentaje de elementos en cada clase. La *frecuencia relativa* de una clase es igual a la parte o proporción de los elementos que pertenecen a cada clase. En un conjunto de datos, en el que hay n observaciones, la frecuencia relativa de cada clase se determina como sigue:

FRECUENCIA RELATIVA

$$\text{Frecuencia relativa de una clase} = \frac{\text{Frecuencia de la clase}}{n} \tag{2.1}$$

La *frecuencia porcentual* de una clase es la frecuencia relativa multiplicada por 100.

Una **distribución de frecuencia relativa** da un resumen tabular de datos en el que se muestra la frecuencia relativa de cada clase. Una **distribución de frecuencia porcentual** da la frecuencia porcentual de los datos de cada clase. En la tabla 2.3 se presenta una distribución de frecuencia relativa y una distribución de frecuencia porcentual de los datos de los refrescos. En esta tabla se observa que la frecuencia relativa de la Coca cola clásica es 19/50 = 0.38, la de la Coca cola de dieta es 8/50 = 0.16, etc. En la distribución de frecuencia porcentual, se muestra que 38% de las ventas fueron de Coca cola clásica, 16% de Coca cola de dieta, etc. También resulta que 38% + 26% + 16% = 80% de las ventas fueron de los tres refrescos que más se venden.

Gráficas de barra y gráficas de pastel

Una **gráfica de barras** o un diagrama de barras, es una gráfica para representar los datos cualitativos de una distribución de frecuencia, de frecuencia relativa o de frecuencia porcentual. En uno de los ejes de la gráfica (por lo general en el horizontal), se especifican las etiquetas empleadas para las clases (categorías). Para el otro eje de la gráfica (el vertical) se usa una escala para

TABLA 2.3 DISTRIBUCIONES DE FRECUENCIA RELATIVA Y FRECUENCIA PORCENTUAL DE LAS VENTAS DE REFRESCOS

Refresco	Frecuencia relativa	Frecuencia porcentual
Coke Classic	0.38	38
Diet Coke	0.16	16
Dr. Pepper	0.10	10
Pepsi	0.26	26
Sprite	0.10	10
Total	1.00	100

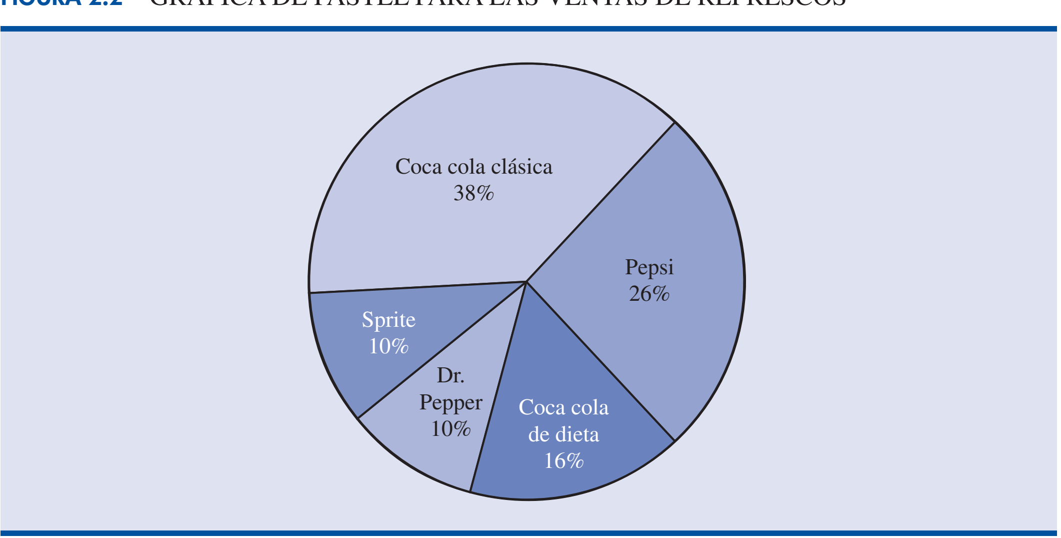
FIGURA 2.1 GRÁFICA DE BARRAS PARA LAS VENTAS DE REFRESCOS

En el control de calidad, las gráficas de barras se usan para identificar las principales causas de problemas. Las graficas se acomodan en orden de alturas descendentes de izquierda a derecha colocando primero la causa de frecuencia más común en primer lugar. A esta gráfica de barras se le llama diagrama de Pareto en honor a su inventor Wilfredo Pareto, un economista italiano.

frecuencia, frecuencia relativa o frecuencia porcentual. Después, empleando un ancho de barra fijo, se dibuja sobre cada etiqueta de las clases una barra que se extiende hasta la frecuencia, frecuencia relativa o frecuencia porcentual de la clase. Cuando se tienen datos cualitativos, las barras deben estar separadas para hacer énfasis en que cada clase está separada. En la figura 2.1 se muestra una gráfica de barras correspondiente a la distribución de frecuencia de las 50 ventas de refrescos. Advierta cómo en esta representación gráfica se observa que Coca cola clásica, Pepsi y Coca cola de dieta son los refrescos preferidos.

La **gráfica de pastel** proporciona otra gráfica para presentar distribuciones de frecuencia relativa y de frecuencia porcentual de datos cualitativos. Para elaborar una gráfica de pastel, primero se dibuja un círculo que representa todos los datos. Después se usa la frecuencia relativa para subdividir el círculo en sectores, o partes, que corresponden a la frecuencia relativa de cada clase. Por ejemplo, como un círculo tiene 360 grados y Coca cola clásica presenta una frecuencia relativa de 0.38, el sector de la gráfica de pastel correspondiente a Coca cola clásica resultará de $0.38(360) = 136.8$ grados. El sector del pastel para Coca cola de dieta constará de

FIGURA 2.2 GRÁFICA DE PASTEL PARA LAS VENTAS DE REFRESCOS



0.16(360) = 57.6 grados. Mediante cálculos semejantes para las demás clases se obtiene la gráfica de pastel de la figura 2.2. Los números que aparecen en cada sector pueden ser frecuencia, frecuencia relativa o frecuencia porcentual.

NOTAS Y COMENTARIOS

1. A menudo el número de clases en una distribución de frecuencia es el mismo que el número de categorías encontradas en los datos, como en los datos de las ventas de refresco en esta sección. Los datos comprenden cinco refrescos y para cada uno se definió una clase en la distribución de frecuencia. Si los datos incluyeran todos los refrescos se requerirían muchas categorías, la mayor parte de las cuales sólo tendrían muy pocas ventas. La mayoría de los profesionistas de la estadística aconsejan que las clases con frecuencia pequeña, se agrupen en una sola clase a la que se le llama "otros". Cualquier clase con 5% o menos se trata de esta manera.

2. La suma de las frecuencias en una distribución de frecuencia es siempre igual al número de observaciones. La suma de las frecuencias relativas en una distribución de frecuencia relativa es siempre igual a 1.00, y la suma de los porcentajes en una distribución de frecuencia porcentual es siempre igual a 100.

Ejercicios

Métodos

1. Como respuesta a una pregunta hay tres alternativas: A, B y C. En una muestra de 120 respuestas, 60 fueron A, 24 B y 36 C. Dé las distribuciones de frecuencia y de frecuencia relativa.
2. Se da una distribución de frecuencia relativa.

Clase	Frecuencia relativa
A	0.22
B	0.18
C	0.40
D	

   a. ¿Cuál es la frecuencia relativa de la clase D?
   b. El tamaño de la muestra es 200. ¿Cuál es la frecuencia de la clase D?
   c. Muestre la distribución de frecuencia.
   d. Dé la distribución de frecuencia porcentual.





3. Un cuestionario proporciona como respuestas 58 Sí, 42 No y 20 ninguna opinión.
   a. En la construcción de una gráfica de pastel, ¿cuántos grados le corresponderán del pastel a la respuesta Sí?
   b. ¿Cuántos grados le corresponderán del pastel a la respuesta No?
   c. Construya una gráfica de pastel.
   d. Construya una gráfica de barras.

Aplicaciones

archivo en CD
TVMedia

4. Los cuatro programas con horario estelar de televisión son *CSI, ER, Everybody Loves Raymond* y *Friends* (*Nielsen Media Research,* 11 de enero de 2004). A continuación se presentan los datos sobre las preferencias de los 50 televidentes de una muestra.

CSI	Friends	CSI	CSI	CSI
CSI	CSI	Raymond	ER	ER
Friends	CSI	ER	Friends	CSI
ER	ER	Friends	CSI	Raymond
CSI	Friends	CSI	CSI	Friends
ER	ER	ER	Friends	Raymond
CSI	Friends	Friends	CSI	Raymond
Friends	Friends	Raymond	Friends	CSI
Raymond	Friends	ER	Friends	CSI
CSI	ER	CSI	Friends	ER

a. ¿Estos datos son cualitativos o cuantitativos?
b. Proporcione las distribuciones de frecuencia y de frecuencia relativa.
c. Construya una gráfica de barras y una gráfica de pastel.
d. De acuerdo con la muestra, ¿qué programa de televisión tiene la mayor audiencia? ¿Cuál es el segundo?

5. Los cinco apellidos más comunes en Estados Unidos, en orden alfabético son, Brown, Davis, Johnson, Jones, Smith y Williams (*The World Almanac, 2006*). Suponga que en una muestra de 50 personas con uno de estos apellidos se obtienen los datos siguientes.



Names

Brown	Williams	Williams	Williams	Brown
Smith	Jones	Smith	Johnson	Smith
Davis	Smith	Brown	Williams	Johnson
Johnson	Smith	Smith	Johnson	Brown
Williams	Davis	Johnson	Williams	Johnson
Williams	Johnson	Jones	Smith	Brown
Johnson	Smith	Smith	Brown	Jones
Jones	Jones	Smith	Smith	Davis
Davis	Jones	Williams	Davis	Smith
Jones	Johnson	Brown	Johnson	Davis


Resuma estos datos construyendo:
a. Distribuciones de frecuencia relativa y porcentual.
b. Una gráfica de barras.
c. Una gráfica de pastel.
d. De acuerdo con estos datos, ¿cuáles son los tres apellidos más comunes?

6. El índice de audiencia de televisión de Nielsen Media Research mide el porcentaje de personas que tienen televisión y que están viendo un determinado programa. El programa de televisión con el mayor índice de audiencia en la historia de la televisión (en Estados Unidos) fue *M*A*S*H Last Episode Special* transmitido el 28 de febrero de 1983. El índice de audiencia de 60.2 indicó que 60.2% de todas las personas que tenían televisión estaban viendo este programa. Nielsen Media Research publicó la lista de los 50 programas de televisión con los mayores índices de audiencia en la historia de la televisión (*The New York Times Almanac*, 2006). Los datos siguientes presentan las cadenas de televisión que produjeron estos 50 programas con mayor índice de audiencia.



Networks

ABC	ABC	ABC	NBC	CBS
ABC	CBS	ABC	ABC	NBC
NBC	NBC	CBS	ABC	NBC
CBS	ABC	CBS	NBC	ABC
CBS	NBC	NBC	CBS	NBC
CBS	CBS	CBS	NBC	NBC
FOX	CBS	CBS	ABC	NBC
ABC	ABC	CBS	NBC	NBC
NBC	CBS	NBC	CBS	CBS
ABC	CBS	ABC	NBC	ABC

a. Con estos datos construya una distribución de frecuencia, una de frecuencia porcentual y una gráfica de barras.

b. ¿Cuál o cuáles cadenas de televisión han presentado los programas de mayor índice de audiencia? Compare los desempeños de ABC, CBS y NBC.



7. Un restaurante de Florida emplea cuestionarios en los que pide a sus clientes que evalúen el servicio, la calidad de los alimentos, los cocteles, los precios y la atmósfera del restaurante. Cada uno de estos puntos se evalúa con una escala de óptimo (O), muy bueno (V), bueno (G), regular (A) y malo (P). Emplee la estadística descriptiva para resumir los datos siguientes respecto a la calidad de los alimentos. ¿Qué piensa acerca de la evaluación de la calidad de los alimentos de este restaurante?

G	O	V	G	A	O	V	O	V	G	O	V	A
V	O	P	V	O	G	A	O	O	O	G	O	V
V	A	G	O	V	P	V	O	O	G	O	O	V
O	G	A	O	V	O	O	G	V	A	G		

8. A continuación se muestran datos de 55 miembros de un equipo de béisbol. Cada observación indica la posición principal que juegan los miembros del equipo: *pitcher* (P), *catcher* (H), primera base (1), segunda base (2), tercera base (3), shortstop (S), left fiel (L), center field (C) y right field (R).

L	P	C	H	2	P	R	1	S	S	1	L	P	R	P
P	P	P	R	C	S	L	R	P	C	C	P	P	R	P
2	3	P	H	L	P	1	C	P	P	P	S	1	L	R
R	1	2	H	S	3	H	2	L	P					

a. Para resumir estos datos use una distribución de frecuencia y otra de frecuencia relativa.
b. ¿Cuál es la posición que ocupan más miembros del equipo?
c. ¿Cuál es la posición que ocupan menos miembros del equipo?
d. ¿Qué posición de campo (L, R, C) es la que juegan más miembros del equipo?
e. Compare las posiciones L, R, y C con las posiciones 1, 2, 3 y S.

9. Cerca del 60% de las empresas pequeñas y medianas son empresas familiares. En un estudio de TEC International se preguntaba al gerente general (CEO, por sus siglas en inglés) cómo había llegado a ese cargo (*The Wall Street Journal*, 16 de diciembre de 2003). Las respuestas fueron que el CEO heredó el negocio, que el CEO formó la empresa o que el CEO estaba contratado por con la empresa. En una muestra de 26 CEOs de empresas familiares, los datos obtenidos acerca de cómo el CEO había llegado a ese puesto fueron los siguientes:



Formó	Formó	Formó	Heredó
Heredó	Formó	Heredó	Formó
Heredó	Formó	Formó	Formó
Formó	Contrató	Contrató	Contrató
Heredó	Heredó	Heredó	Formó
Formó	Formó	Formó	Contrató
Formó	Heredó		

a. Dé una distribución de frecuencias.
b. Dé una distribución de frecuencias porcentuales.
c. Presente una gráfica de barras.
d. ¿Qué porcentaje de los CEOs de empresas familiares llegaron a ese puesto por heredar la empresa? ¿Cuál es la razón principal por la que una persona llega al puesto de CEO en una empresa familiar?

10. Netflix, Inc., de San José California, renta, por correo, más de 50 000 títulos de DVD. Los clientes ordenan en línea los DVDs que deseen ver. Antes de ordenar un DVD, el cliente puede ver una descripción del mismo y, si así lo desea, un resumen de las evaluaciones del mismo. Netflix emplea un sistema de evaluación de cinco estrellas que tienen el significado siguiente:

1 estrella	Me disgustó
2 estrellas	No me disgustó
3 estrellas	Me gustó
4 estrellas	Me gustó mucho
5 estrellas	Me fascinó

Dieciocho críticos, entre los que se encontraban Roger Ebert de *Chicago Sun Times* y Ty Burr de *Boston Globe,* proporcionaron evaluaciones en Hispanoamérica de la película *Batman inicia* (Netflix.com, 1 de marzo de 2006). Las evaluaciones fueron las siguientes:

4, 2, 5, 2, 4, 3, 3, 4, 4, 3, 4, 4, 4, 2, 4, 4, 5, 4

a. Diga por qué son cualitativos estos datos.
b. Dé una distribución de frecuencias y una distribución de frecuencia relativa.
c. Dé una gráfica de barras.
d. Haga un comentario sobre las evaluaciones que dieron los críticos a esta película.



Resumen de datos cuantitativos

Distribución de frecuencia

TABLA 2.4 AUDITORÍA ANUAL (DÍAS DE DURACIÓN)

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

Como se definió en la sección 2.1, una distribución de frecuencia es un resumen de datos tabular que presenta el número de elementos (frecuencia) en cada una de las clases disyuntas. Esta definición es válida tanto para datos cualitativos como cuantitativos. Sin embargo, cuando se trata de datos cuantitativos se debe tener más cuidado al definir las clases disyuntas que se van a usar en la distribución de frecuencia.

Considere, por ejemplo, los datos cuantitativos de la tabla 2.4. En esta tabla se presenta la duración en días de una muestra de auditorías de fin de año de 20 clientes de una empresa pequeña de contadores públicos. Los tres pasos necesarios para definir las clases de una distribución de frecuencia con datos cuantitativos son

1. Determinar el número de clases disyuntas.
2. Determinar el ancho de cada clase
3. Determinar los límites de clase.



Audit

Se mostrarán estos pasos elaborando una distribución de frecuencia con los datos de la tabla 2.4.

Número de clases Las clases se forman especificando los intervalos que se usarán para agrupar los datos. Se recomienda emplear entre 5 y 20 clases. Cuando los datos son pocos, cinco o seis clases bastan para resumirlos. Si son muchos, se suele requerir más clases. La idea es tener las clases suficientes para que se muestre la variación en los datos, pero no deben ser demasiadas si algunas de ellas contienen sólo unos cuantos datos. Como el número de datos en la tabla 2.4 es relativamente pequeña ($n = 20$), se decide elaborar una distribución de frecuencia con cinco clases.

Hacer las clases de una misma amplitud reduce la posibilidad de que los usuarios hagan interpretaciones inapropiadas.

Ancho de clase El segundo paso al construir una distribución de frecuencia para datos cuantitativos es elegir el ancho de las clases. Como regla general es recomendable que el ancho sea el mismo para todas las clases. Así, el ancho y el número de clases no son decisiones independientes. Entre mayor sea el número de clases menor es el ancho de las clases y viceversa. Para determinar el ancho de clase apropiada se empieza por identificar el mayor y el menor de los valores de los datos. Después, usando el número de clases deseado, se emplea la expresión siguiente para determinar el ancho aproximada de clase.

$$\text{Ancho aproximada de clase} = \frac{\text{Valor mayor en los datos} - \text{Valor menor en los datos}}{\text{Número de clase}} \quad \textbf{(2.2)}$$

El ancho aproximado de clase que se obtiene con la ecuación (2.2) se redondea a un valor más adecuado de acuerdo con las preferencias de la persona que elabora la distribución de frecuencia. Por ejemplo, si el ancho de clase aproximado es 9.28, se redondea a 10 porque 10 es un ancho de clase más adecuado para la presentación de la distribución de la frecuencia.

En los datos sobre las duraciones de las auditorías de fin de año el valor mayor en los datos es 33 y el valor menor es 12. Como se ha decidido resumir los datos en cinco clases, empleando

la ecuación (2.2) el ancho aproximado de clase que se obtiene es (33 − 12)/5 = 4.2. Por tanto, al redondear, en la distribución de frecuencia se usa como ancho de clase cinco días.

No hay una distribución de frecuencia que sea la mejor para un conjunto de datos. Distintas personas elaboran diferentes, pero igual de aceptables, distribuciones de frecuencia para un conjunto de datos dado. El objetivo es hacer notar el agrupamiento y la variación natural de los datos.

En la práctica el número de clases y su ancho adecuado se determinan por prueba y error. Una vez que se elige una determinado número de clases, se emplea la ecuación 2.2 para determinar el ancho aproximado de clase. El proceso se repite con distintos números de clases. El analista determina la combinación de número y ancho de clases que le proporciona la mejor distribución de frecuencia para resumir los datos.

En el caso de los datos de la tabla 2.4, una vez que se ha decidido emplear cinco clases, cada una con ancho de cinco días, el paso siguiente es especificar los límites de cada clase.

Límites de clase Los límites de clase deben elegirse de manera que cada dato pertenezca a una y sólo una de las clases. El *límite de clase inferior* indica el menor valor de los datos a que pertenece esa clase. El *límite de clase superior* indica el mayor valor de los datos a que pertenece esa clase. Al elaborar distribuciones de frecuencia para datos cualitativos, no es necesario especificar límites de clase porque cada dato corresponde de manera natural a una de las clases disyuntas. Pero con datos cuantitativos, como la duración de las auditorías de la tabla 2.4, los límites de clase son necesarios para determinar dónde colocar cada dato.

TABLA 2.5

DISTRIBUCIÓN DE FRECUENCIA DE LAS AUDITORÍAS

Duración de las auditorías (días)	Frecuencia
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

Mediante los datos de la duración de las auditorías de la tabla 2.4, se elige 10 días como límite inferior y 14 como límite superior de la primera clase. En la tabla 2.5, esta clase se denota como 10–14. El valor menor, 12 (de la tabla), pertenece a la clase 10–14. Después se elige 15 días como límite inferior y 19 como límite superior de la clase siguiente. Así, se continúan definiendo los límites inferior y superior de las clases hasta tener las cinco clases: 10–14, 15–19, 20–24, 25–29 y 30–34. El valor mayor en los datos, 33, pertenece a la clase 33–34. Las diferencias entre los límites inferiores de clase de clases adyacentes es el ancho de clase. Con los dos primeros límites inferiores de clase, 10 y 15, se ve que el ancho de clase es 15 – 10 = 5

Una vez determinados números, ancho y límites de las clases, la distribución de frecuencia se obtiene contando el número de datos que corresponden a cada clase. Por ejemplo, en la tabla 2.4 se observa que hay cuatro valores, 12, 14, 14 y 13, que pertenecen a la clase 10-14. Por tanto, la frecuencia de la clase 10–14 es 4. Al continuar con este proceso de conteo para las clases 15–19, 20–24, 25–29 y 30–34 se obtiene la distribución de frecuencia que se muestra en la tabla 2.5. En esta distribución de frecuencia se observa lo siguiente:

1. Las duraciones de las auditorías que se presentan con más frecuencia son de la clase 15-19 días. Ocho de las 20 auditorías caen en esta clase.
2. Sólo una auditoría requirió 30 o más días.

También se obtienen otras conclusiones, dependiendo de los intereses de quien observa la distribución de frecuencia. La utilidad de una distribución de frecuencia es que proporciona claridad acerca de los datos, la cual no es fácil de obtener con la forma desorganizada de éstos.

Punto medio de clase En algunas aplicaciones se desea conocer el punto medio de las clases de una distribución de frecuencia de datos cuantitativos. El **punto medio de clase** es el valor que queda a la mitad entre el límite inferior y el límite superior de la clase. En el caso de las duraciones de las auditorías, los cinco puntos medios de clase son 12, 17, 22, 27 y 32.

Distribuciones de frecuencia relativa y de frecuencia porcentual

Las distribuciones de frecuencia relativa y de frecuencia porcentual para datos cuantitativos se definen de la misma forma que para datos cualitativos. Primero debe recordar que la frecuencia relativa es el cociente, respecto al total de observaciones, de las observaciones que pertenecen a una clase. Si el número de observaciones es n,

$$\text{Frecuencia relativa de la clase} = \frac{\text{Frecuencia de la clase}}{n}$$

La frecuencia porcentual de una clase es la frecuencia relativa multiplicada por 100.

Con base en la frecuencia de las clases de la tabla 2.5 y dado que $n = 20$, en la tabla 2.6 se muestran las distribuciones de frecuencia relativa y de frecuencia porcentual de los datos de las

TABLA 2.6 DISTRIBUCIONES DE FRECUENCIA RELATIVA Y DE FRECUENCIA PORCENTUAL CON LOS DATOS DE LAS DURACIONES DE LAS AUDITORÍAS

Duración de las auditorías (días)	Frecuencia relativa	Frecuencia porcentual
10–14	0.20	20
15–19	0.40	40
20–24	0.25	25
25–29	0.10	10
30–34	0.05	5
Total	1.00	100

duraciones de las auditorías. Observe que 0.40 de las auditorías, o 40%, necesitaron entre 15 y 19 días. Sólo 0.05%, o 5%, requirió 30 o más días. De nuevo, hay más interpretaciones o ideas que se obtienen de la tabla 2.6.

Gráficas de puntos

Uno de los más sencillos resúmenes gráficos de datos son las **gráficas de puntos**. En el eje horizontal se presenta el intervalo de los datos. Cada dato se representa por un punto colocado sobre este eje. La figura 2.3 es la gráfica de puntos de los datos de la tabla 2.4. Los tres puntos que se encuentran sobre el 18 del eje horizontal indican que hubo tres auditorías de 18 días. Las gráficas de puntos muestran los detalles de los datos y son útiles para comparar la distribución de los datos de dos o más variables.

Histograma

Una presentación gráfica usual para datos cuantitativos es el **histograma**. Esta gráfica se hace con datos previamente resumidos mediante una distribución de frecuencia, de frecuencia relativa o de frecuencia porcentual. Un histograma se construye colocando la variable de interés en el eje horizontal y la frecuencia, la frecuencia relativa o la frecuencia porcentual en el eje vertical. La frecuencia, frecuencia relativa o frecuencia porcentual de cada clase se indica dibujando un rectángulo cuya base está determinada por los límites de clase sobre el eje horizontal y cuya altura es la frecuencia, la frecuencia relativa o la frecuencia porcentual correspondiente.

La figura 2.4 es un histograma de las duraciones de las auditorías. Observe que la clase con mayor frecuencia se indica mediante el rectángulo que se encuentra sobre la clase 15–19 días. La altura del rectángulo muestra que la frecuencia de esta clase es 8. Un histograma de las distribuciones de frecuencia relativa o porcentual de estos datos se ve exactamente igual que el histograma de la figura 2.4, excepto que en el eje vertical se colocan los valores de frecuencia relativa o porcentual.

Como se muestra en la figura 2.4, los rectángulos adyacentes de un histograma se tocan uno a otro. A diferencia de las gráficas de barras, en un histograma no hay una separación natural en-



FIGURA 2.3 GRÁFICA DE PUNTOS PARA LOS DATOS DE LAS DURACIONES DE LAS AUDITORÍAS

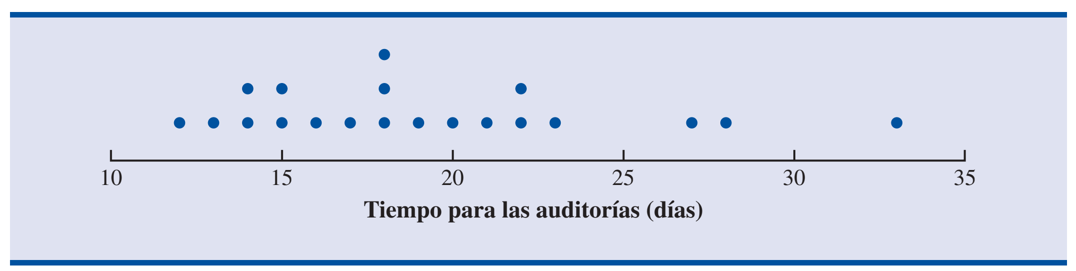
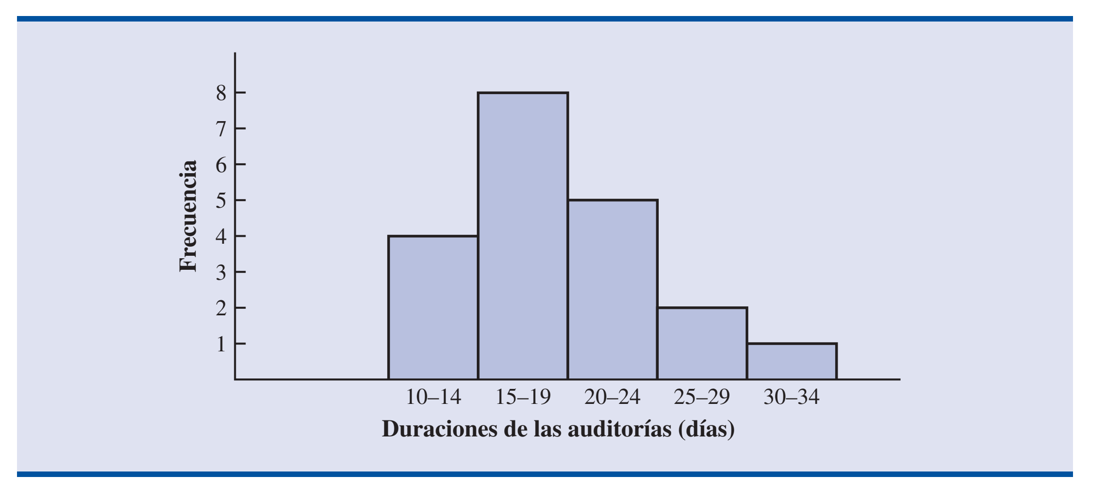


FIGURA 2.4 HISTOGRAMA DE LOS DATOS DE LAS DURACIONES DE LAS AUDITORÍAS



tre los rectángulos de clases adyacentes. Este formato es el usual para histogramas. Como las clases de las duraciones de las auditorías son 10–14, 15–19, 20–24, 25–29 y 30–34 parecería que se necesitara una unidad de espacio entre las clases, de 14 a 15, de 19 a 20, de 24 a 25 y de 29 a 30. Cuando se construye un histograma se eliminan estos espacios. Eliminar los espacios entre las clases del histograma de las duraciones de las auditorías sirve para indicar que todos los valores entre el límite inferior de la primera clase y el superior de la última son posibles.

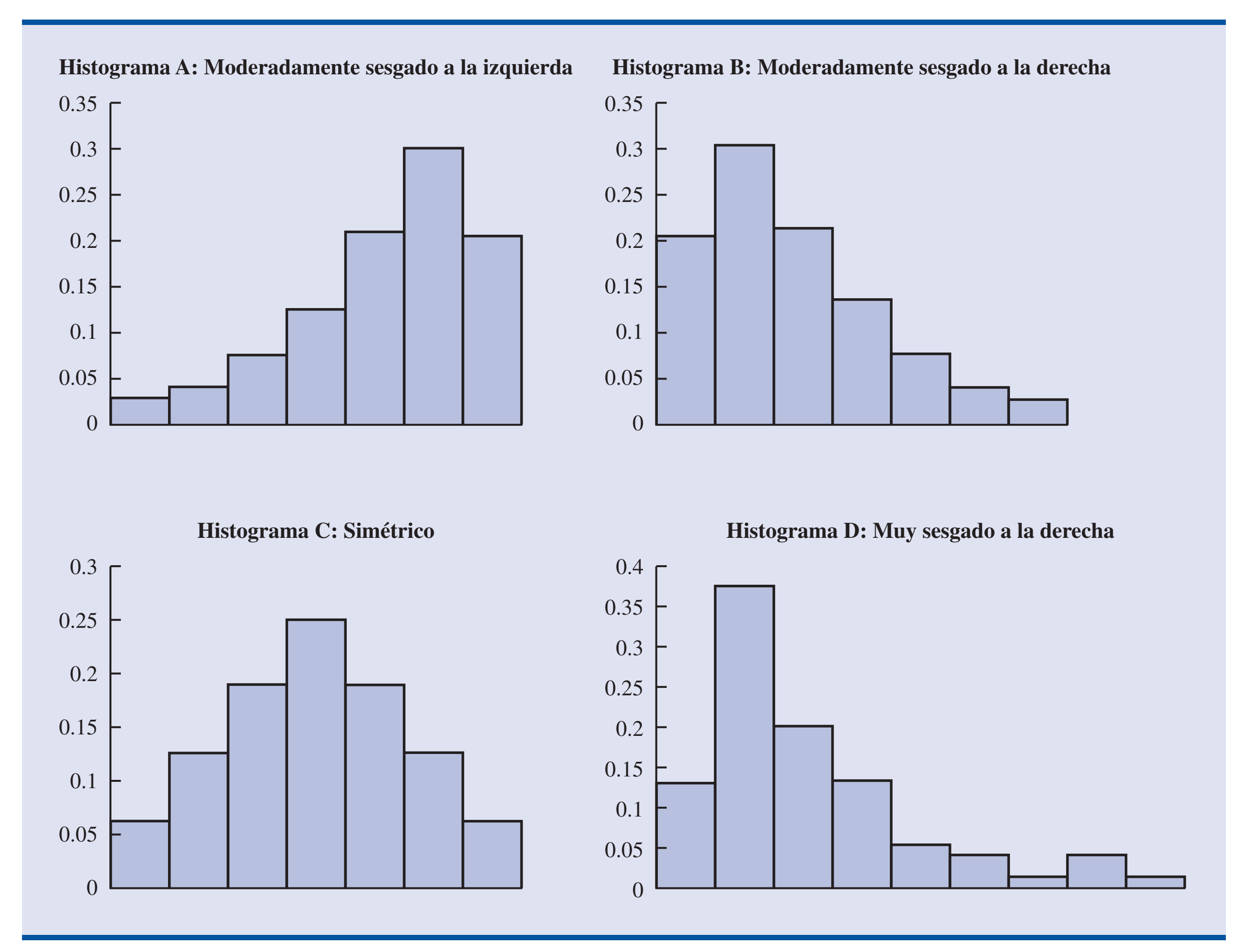
Uno de los usos más importantes de un histograma es proveer información acerca de la forma de la distribución. En la figura 2.5 se muestran cuatro histogramas construidos a partir de distribuciones de frecuencia relativa. En el histograma A se muestra un conjunto de datos moderadamente sesgado a la izquierda. Se dice que un histograma es sesgado a la izquierda si su cola se extiende más hacia la izquierda. Dichos histogramas son típicos para calificaciones: no hay calificaciones mayores a 100%, la mayor parte están arriba de 70% y sólo hay unas cuantas bajas. En el histograma B se muestra un conjunto de datos moderadamente sesgado a la derecha. Un histograma está sesgado a la derecha si su cola se extiende más hacia la derecha. Ejemplos de este tipo de histogramas son los datos de los precios de las casas; unas cuantas casas caras crean el sesgo a la derecha.

En C se observa un histograma simétrico. En éste la cola izquierda es la imagen de la cola derecha. Los histogramas de datos para aplicaciones nunca son perfectamente simétricos, pero en muchas aplicaciones suelen ser más o menos simétricos. En D se observa un histograma muy sesgado a la derecha. Éste se elaboró con datos sobre la cantidad de compras a lo largo de un día en una tienda de ropa para mujeres. Los datos de aplicaciones de negocios o economía suelen conducir a histogramas sesgados a la derecha. Por ejemplo datos de los precios de las casas, de los salarios, de las cantidades de las compras, etc., suelen dar histogramas sesgados a la derecha.

Distribuciones acumuladas

Una variación de las distribuciones de frecuencia que proporcionan otro resumen tabular de datos cuantitativos es la **distribución de frecuencia acumulada**. La distribución de frecuencia acumulada usa la cantidad, las amplitudes y los límites de las clases de la distribución de frecuencia. Sin embargo, en lugar de mostrar la frecuencia de cada clase, la distribución de frecuencia acumulada muestra la cantidad de datos que tienen un valor *menor o igual* al límite superior de cada clase. Las primeras dos columnas de la tabla 2.7 corresponden a la distribución de frecuencia acumulada de los datos de las duraciones de las auditorías.

FIGURA 2.5 HISTOGRAMAS CON DISTINTOS TIPOS DE SESGO



Para entender cómo se determina la frecuencia acumulada, considere la clase que dice "menor o igual que 24". La frecuencia acumulada en esta clase es simplemente la suma de la frecuencia de todas las clases en que los valores de los datos son menores o iguales que 24. En la distribución de frecuencia de la tabla 2.5 la suma de las frecuencias para las clases 10–14, 15–29 y 20–24 indica que los datos cuyos valores son menores o iguales que 24 son 4 + 8 + 5 = 17. Por lo tanto, en esta clase la frecuencia acumulada es 17. Además, en la ditribución de frecuen-

TABLA 2.7 DISTRIBUCIONES DE FRECUENCIA ACUMULADA, FRECUENCIA RELATIVA ACUMULADA Y FRECUENCIA PORCENTUAL ACUMULADA

Duración de la auditoría en días	Frecuencia acumulada	Frecuencia relativa acumulada	Frecuencia porcentual acumulada
Menor o igual que 14	4	0.20	20
Menor o igual que 19	12	0.60	60
Menor o igual que 24	17	0.85	85
Menor o igual que 29	19	0.95	95
Menor o igual que 34	20	1.00	100

cias acumuladas de la tabla 2.7 se observa que cuatro auditorías duraron 14 días o menos y que 19 auditorías duraron 29 días o menos.

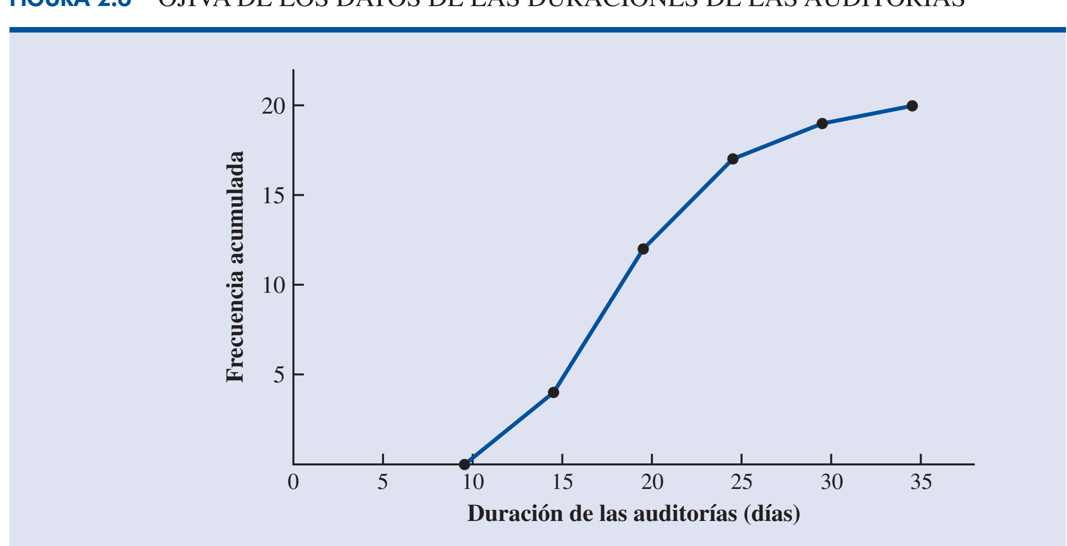
Por último, se tiene que la **distribución de frecuencias relativas acumuladas** indica la proporción de todos los datos que tienen valores menores o iguales al límite superior de cada clase, y la **distribución de frecuencias porcentuales acumuladas** indica el porcentaje de todos los datos que tienen valores menores o iguales al límite superior de cada clase. La distribución de frecuencias relativas acumuladas se calcula ya sea sumando las frecuencias relativas que aparecen en la distribución de frecuencias relativas o dividiendo la frecuencia acumulada entre la cantidad total de datos. Empleando el último método, las frecuencias relativas acumuladas que aparecen en la columna 3 de la tabla 2.7 se obtienen dividiendo las frecuencias acumuladas de la columna 2 entre la cantidad total de datos ($n = 20$). Las frecuencias porcentuales acumuladas se obtienen multiplicando las frecuencias relativas por 100. Estas distribuciones de frecuencias acumuladas relativas y porcentuales indican que 0.85 o el 85% de las auditorías se realizaron en 24 días o menos, 0.95 o 95% de las auditorías se realizaron en 29 días o menos, etcétera.

Ojiva

La gráfica de una distribución acumulada, llamada **ojiva**, es una gráfica que muestra los valores de los datos en el eje horizontal y las frecuencias acumuladas, las frecuencias relativas acumuladas o las frecuencias porcentuales acumuladas en el eje vertical. En la figura 2.6 se muestra una ojiva correspondiente a las frecuencias acumuladas de las duraciones de las auditorías.

La ojiva se construye al graficar cada uno de los puntos correspondientes a la frecuencia acumulada de las clases. Como las clases de las duraciones de las auditorías son 10–14, 15–19, 20–24, etc., hay huecos de una unidad entre 14 y 15, 19 y 20, etc. Estos huecos se eliminan al graficar puntos a la mitad entre los dos límites de clase. Así, para la clase 10–14 se usa 14.5, para la clase 15–19 se usa 19.5 y así en lo sucesivo. En la ojiva de la figura 2.6 la clase "menor o igual que 14" cuya frecuencia acumulada es 4 se grafica mediante el punto que se localiza a 14.5 unidades sobre el eje horizontal y a 4 unidades sobre el vertical. La clase "menor o igual que 19" cuya frecuencia acumulada es 12 se representa por un punto que se encuentra a 19.5 unidades sobre el eje horizontal y 12 unidades sobre el vertical. Observe que en el extremo izquierdo de la ojiva se ha graficado un punto más. Este punto inicia la ojiva mostrando que en los datos no hay valores que se encuentren abajo de la clase 10–14. Este punto se encuentra a 9.5 unidades sobre el eje horizontal y a 0 unidades sobre el vertical. Para terminar los puntos graficados se conectan mediante líneas rectas.

FIGURA 2.6 OJIVA DE LOS DATOS DE LAS DURACIONES DE LAS AUDITORÍAS



NOTAS Y COMENTARIOS

1. Una gráfica de barras y un histograma son en esencia lo mismo; ambas son representaciones gráficas de una distribución de frecuencia. Un histograma es sólo una gráfica de barras sin separación entre las barras. Para algunos datos cuantitativos discretos, también se puede tener separación entre las barras. Considere por ejemplo, el número de materias en que está inscrito un estudiante universitario. Los datos sólo tienen valores enteros. No hay valores intermedios como 1.5, 2.73, etc. Sin embargo cuando se tienen datos cuantitativos continuos, como en las auditorías, no es apropiado tener separación entre las barras.
2. Los valores adecuados para los límites de clase cuando se tienen datos cuantitativos depende del nivel de precisión de los datos. Por ejemplo, en el caso de los datos de la tabla 2.4, sobre la duración de las auditorías, los límites usados fueron números enteros. Si los datos hubieran estado redondeados a la décima de día más cercana (es decir, 12.3, 14.4, etc.), entonces los límites se hubieran dado con décimas de día. La primera clase, por ejemplo, hubiera sido de 10.0 a 14.9. Si los datos se hubieran registrado hasta la centésima de día más cercana (es decir, 12.34, 14.45, etc.), los límites se hubieran dado con centésimas de días. Por ejemplo la primera clase hubiera sido de 10.00–14.99.
3. Una clase *abierta* sólo necesita el límite inferior de la clase o el límite superior de la clase. Por ejemplo, suponga que en los datos de la tabla 2.4 sobre las duraciones de las auditorías dos de éstas hubieran durado 58 y 65 días. En lugar de haber seguido con clases de amplitud 5 de 35–39, de 40–44, de 45 a 49, etc., podría haber simplificado la distribución de frecuencia mediante una clase abierta de "35 o más". La frecuencia de esta clase habría sido 2. La mayor parte de las clases abiertas aparecen en el extremo superior de la distribución. Algunas veces se encuentran clases abiertas en el extremo inferior y rara vez están en ambos extremos.
4. En una distribución de frecuencia acumulada, la última frecuencia siempre es igual al número total de observaciones. En una distribución de frecuencia relativa acumulada la última frecuencia siempre es igual a 1.00 y en una distribución de frecuencia porcentual acumulada la última frecuencia es siempre 100.

Ejercicios

Métodos

11. Considere los datos siguientes.



14	21	23	21	16
19	22	25	16	16
24	24	25	19	16
19	18	19	21	12
16	17	18	23	25
20	23	16	20	19
24	26	15	22	24
20	22	24	22	20

a. Elabore una distribución de frecuencia usando las clases 12–14, 15–17, 18–20, 21–23 y 24–26.
b. Elabore una distribución de frecuencia relativa y una de frecuencia porcentual usando las clases del inciso a.



12. Considere la distribución de frecuencia siguiente.

Clases	Frecuencia
10–19	10
20–29	14
30–39	17
40–49	7
50–59	2

Construya una distribución de frecuencia acumulada y otra de frecuencia relativa acumulada.

13. Con los datos del ejercicio 12 elabore un histograma y una ojiva.

14. Considere los datos siguientes.

8.9	10.2	11.5	7.8	10.0	12.2	13.5	14.1	10.0	12.2
6.8	9.5	11.5	11.2	14.9	7.5	10.0	6.0	15.8	11.5

a. Construya un diagrama de punto.
b. Elabore una distribución de frecuencia.
c. Construya una distribución de frecuencia porcentual.

Aplicaciones



15. El personal de un consultorio analiza los tiempos de espera de los pacientes que requieren servicio de emergencia. Los datos siguientes son los tiempos de espera en minutos recolectados a lo largo de un mes.

2 5 10 12 4 4 5 17 11 8 9 8 12 21 6 8 7 13 18 3

Con las clases 0–4, 5–9, etcétera.
a. Muestre la distribución de la frecuencia.
b. Exprese la distribución de la frecuencia relativa.
c. Muestre la distribución de frecuencia acumulada.
d. Presente la distribución de frecuencia relativa acumulada.
e. ¿Cuál es la proporción de los pacientes que requieren servicio de emergencia y esperan 9 minutos o menos?

16. Considere las dos distribuciones de frecuencias siguientes. La primera distribución de frecuencia proporciona el ingreso anual bruto ajustado de Estados Unidos (Internal Revenue Service, marzo 2003). La segunda distribución de frecuencia muestra las calificaciones de exámenes de un grupo de estudiantes universitarios en un curso de estadística.

Ingreso (en miles de $)	**Frecuencia (en millones)**	**Calificaciones de examen**	**Frecuencia**
0–24	60	20–29	2
25–49	33	30–39	5
50–74	20	40–49	6
75–99	6	50–59	13
100–124	4	60–69	32
125–149	2	70–79	78
150–174	1	80–89	43
175–199	1	90–99	21
Total	127	Total	200

a. Con los datos del ingreso anual elabore un histograma. ¿Qué evidencia de sesgo observa? ¿Es razonable este sesgo? Explique.
b. Con los datos de las calificaciones elabore un histograma. ¿Qué evidencia de sesgo observa? Explique.
c. Con los datos del ejercicio 11 elabore un histograma. ¿Qué evidencia de sesgo observa? ¿Cuál es la forma general de la distribución?

17. ¿Cuál es el precio típico de las acciones de las 30 empresas del promedio industrial Dow Jones? Los datos siguientes son los precios de las acciones, al dólar más cercano, en enero de 2006 (*The Wall Street Journal,* 16 de enero de 2006).



Empresa	$/Acción	Empresa	$/Acción
AIG	70	Home Depot	42
Alcoa	29	Honeywell	37
Altria Group	76	IBM	83
American Express	53	Intel	26
AT&T	25	Johnson & Johnson	62
Boeing	69	JPMorgan Chase	40
Caterpillar	62	McDonald's	35
Citigroup	49	Merck	33
Coca-Cola	41	Microsoft	27
Disney	26	3M	78
DuPont	40	Pfizer	25
ExxonMobil	61	Procter & Gamble	59
General Electric	35	United Technologies	56
General Motors	20	Verizon	32
Hewlett-Packard	32	Wal-Mart	45

a. Con estos datos elabore una distribución de frecuencia.
b. Con estos datos elabore un histograma. Interprete el histograma, presente un análisis de la forma general del histograma, el precio medio de cada intervalo de acciones, el precio más frecuente por intervalo de acciones, los precios más alto y más bajo por acción.
c. ¿Cuáles son las acciones que tienen el precio más alto y el más bajo?
d. Use *The Wall Street Journal* para encontrar los precios actuales por acción de estas empresas. Elabore un histograma con estos datos y discuta los cambios en comparación con enero de 2006.

18. NRF/BIG proporciona los resultados de una investigación sobre las cantidades que gastan en vacaciones los consumidores (*USA Today,* 20 de diciembre de 2005). Los datos siguientes son las cantidades gastadas en vacaciones por los 25 consumidores de una muestra.



1200	850	740	590	340
450	890	260	610	350
1780	180	850	2050	770
800	1090	510	520	220
1450	280	1120	200	350

a. ¿Cuál es la menor cantidad gastada en vacaciones? ¿Cuál la mayor?
b. Use $250 como amplitud de clase para elaborar con estos datos una distribución de frecuencia y una distribución de frecuencia porcentual.
c. Elabore un histograma y comente la forma de la distribución.
d. ¿Qué observaciones le permiten hacer las cantidades gastadas en vacaciones?

19. El correo no deseado afecta la productividad de los oficinistas. Se hizo una investigación con oficinistas para determinar la cantidad de tiempo por día que pierden en estos correos no deseados. Los datos siguientes corresponden a los tiempos en minutos perdidos por día observados en una muestra.

2	4	8	4
8	1	2	32
12	1	5	7
5	5	3	4
24	19	4	14

Resuma estos datos construyendo:
a. Una distribución de frecuencia (con las clases 1–5, 6–10, 11–15, 16–20, etc.)
b. Una distribución de frecuencia relativa
c. Una distribución de frecuencia acumulada.

d. Una distribución de frecuencia relativa acumulada.
e. Una ojiva.
f. ¿Qué porcentaje de los oficinistas pierde 5 minutos o menos en revisar el correo no deseado? ¿Qué porcentaje pierde más de 10 minutos por día en esto?

20. A continuación se presentan las 20 mejores giras de concierto y el precio promedio del costo de sus entradas en Estados Unidos. Esta lista se basa en datos proporcionados por los promotores y administradores de los locales a la publicación *Pollstar* (*Associated Press*, 21 de noviembre de 2003).



Gira de conciertos	Precio de la entrada	Gira de conciertos	Precio de la entrada
Bruce Springsteen	$72.40	Toby Keith	$37.76
Dave Matthews Band	44.11	James Taylor	44.93
Aerosmith/KISS	69.52	Alabama	40.83
Shania Twain	61.80	Harper/Johnson	33.70
Fleetwood Mac	78.34	50 Cent	38.89
Radiohead	39.50	Steely Dan	36.38
Cher	64.47	Red Hot Chili Peppers	56.82
Counting Crows	36.48	R.E.M.	46.16
Timberlake/Aguilera	74.43	American Idols Live	39.11
Mana	46.48	Mariah Carey	56.08

Resuma los datos construyendo:
a. Una distribución de frecuencia y una distribución de frecuencia porcentual.
b. Un histograma.
c. ¿Qué concierto tiene el precio promedio más alto? ¿Qué concierto tiene el precio promedio menos caro?
d. Haga un comentario sobre qué indican los datos acerca de los precios promedio de las mejores giras de concierto.

21. *Nielsen Home Technology Report* informa sobre la tecnología en el hogar y su uso. Los datos siguientes son las horas de uso de computadora por semana en una muestra de 50 personas.



4.1	1.5	10.4	5.9	3.4	5.7	1.6	6.1	3.0	3.7
3.1	4.8	2.0	14.8	5.4	4.2	3.9	4.1	11.1	3.5
4.1	4.1	8.8	5.6	4.3	3.3	7.1	10.3	6.2	7.6
10.8	2.8	9.5	12.9	12.1	0.7	4.0	9.2	4.4	5.7
7.2	6.1	5.7	5.9	4.7	3.9	3.7	3.1	6.1	3.1

Resuma estos datos construyendo:
a. Una distribución de frecuencia (como ancho de clase use tres horas).
b. Una distribución de frecuencia relativa.
c. Un histograma.
d. Una ojiva.
e. Haga un comentario sobre lo que indican los datos respecto al uso de la computadora en el hogar.



Análisis exploratorio de datos: el diagrama de tallo y hojas

Las técnicas del **análisis exploratorio de datos** emplean aritmética sencilla y gráficas fáciles de dibujar útiles para resumir datos. La técnica conocida como **diagrama de tallo y hojas** muestra en forma simultánea el orden jerárquico y la forma de un conjunto de datos.

TABLA 2.8 NÚMERO DE PREGUNTAS CONTESTADAS CORRECTAMENTE EN UN EXAMEN DE APTITUDES



ApTest

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

Para ilustrar el uso de los diagramas de tallo y hojas, considere la tabla 2.8. Estos datos son el resultado de un examen de aptitudes con 150 preguntas presentado por 50 personas que aspiraban a un puesto en una empresa. Los datos indican el número de respuestas correctas por examen.

Para elaborar un diagrama de tallo y hoja inicie acomodando los primeros dígitos de cada uno de los datos a la izquierda de una línea vertical. A la derecha de la línea vertical se anota el último dígito de cada dato. Con base en el primer renglón de la tabla 2.8 (112, 72, 69, 97 y 107), los primeros cinco datos al elaborar el diagrama de tallo y hojas serían los siguientes:

6	9
7	2
8	
9	7
10	7
11	2
12	
13	
14	

Por ejemplo, para el dato 112, se observa que los primeros dígitos, 11, se encuentran a la izquierda de la línea y el último dato, 2, a la derecha. De manera similar, el primer dígito, 7, del dato 72 se encuentra a la izquierda de la línea y el 2 a la derecha. Si continúa colocando el último dígito de cada dato en el renglón correspondiente a sus primeros dígitos obtiene:

6	9 8
7	2 3 6 3 6 5
8	6 2 3 1 1 0 4 5
9	7 2 2 6 2 1 5 8 8 5 4
10	7 4 8 0 2 6 6 0 6
11	2 8 5 9 3 5 9
12	6 8 7 4
13	2 4
14	1

Una vez organizados los datos de esta manera, ordenar los datos de cada renglón de menor a mayor es sencillo. Entonces obtiene el diagrama de tallo y hojas que se muestra aquí.

6	8 9
7	2 3 3 5 6 6
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

Los números a la izquierda de la línea vertical (6, 7, 8, 9, 10, 11, 12, 13 y 14) forman el *tallo,* y cada dígito a la derecha de la línea vertical es una *hoja*. Por ejemplo, considere el primer renglón que tiene como tallo el 6 y como hojas 8 y 9.

6 | 8 9

Este renglón indica que hay dos datos que tienen como primer dígito el seis. Las hojas indican que estos datos son 68 y 69. De manera similar, el segundo renglón

7 | 2 3 3 5 6 6

indica que hay seis datos que tienen como primer dígito el 7. Las hojas indican que estos datos son 72, 73, 73, 75, 76 y 76.

Para atender a la forma del diagrama de tallo y hojas, se usan rectángulos que contienen las hojas de cada tallo; con esto se obtiene lo siguiente.

6	8 9
7	2 3 3 5 6 6
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

Al rotar la página sobre su costado en contra de las manecillas del reloj se obtiene una imagen de los datos que es parecida a un histograma y en el que las clases son 60–69, 70–79, 80–89, etcétera.

Aunque el diagrama de tallo y hojas parece proporcionar la misma información que un histograma, tiene dos ventajas fundamentales.

1. El diagrama de tallo y hojas es más fácil de construir a mano.
2. En cada intervalo de clase proporciona más información que un histograma debido a que el tallo y la hoja proporcionan el dato.

Así como para una distribución de frecuencia o para un histograma no hay un determinado número de clases, tampoco para el diagrama de tallo y hojas hay un número determinado de renglones a tallos. Si piensa que el diagrama de tallo y hojas original condensa demasiado los datos, es fácil expandirlo empleando dos o más tallos por cada primer dígito. Por ejemplo, para usar

En un diagrama expandido de tallo y hojas, siempre que un tallo aparece dos veces, al primero le corresponden las hojas 0–4 y al segundo las hojas 5–9.

dos tallos por cada primer dígito se ponen todos los datos que terminen en 0, 1, 2, 3 o 4 en un renglón y todos los datos que terminen en 5, 6, 7, 8 o 9 en otro. Este método se ilustra en el siguiente diagrama expandido de tallo y hojas.

6	8 9
7	2 3 3
7	5 6 6
8	0 1 1 2 3 4
8	5 6
9	1 2 2 2 4
9	5 5 6 7 8 8
10	0 0 2 4
10	6 6 6 7 8
11	2 3
11	5 5 8 9 9
12	4
12	6 7 8
13	2 4
13	
14	1

Observe que las hojas de los datos 72, 73 y 73 pertenecen al intervalo 0–4 y aparecen con el primer tallo que tiene el valor 7. Las hojas de los valores 75, 76 y 76 pertenecen al intervalo 5–9 y aparecen con el segundo tallo que tiene el valor 7. Este diagrama expandido de tallo y hojas es semejante a una distribución con los intervalos 65–69, 70–74, 75–79, etcétera.

El ejemplo anterior muestra un diagrama de tallo y hojas con datos de hasta tres dígitos. Estos diagramas también se elaboran con datos de más de tres dígitos. Por ejemplo, considere los datos siguientes sobre el número de hamburguesas vendidas en un restaurante de comida rápida en cada una de 15 semanas.

1565	1852	1644	1766	1888	1912	2044	1812
1790	1679	2008	1852	1967	1954	1733	

A continuación se presenta un diagrama de tallo y hojas de estos datos.

Unidad de hoja = 10

15	6
16	4 7
17	3 6 9
18	1 5 5 8
19	1 5 6
20	0 4

En un diagrama de tallo y hojas se usa un solo dígito para definir cada hoja. La unidad de hoja indica por qué número debe multiplicar los números del tallo y la hoja para aproximar el dato original. Las unidades de hoja son 100, 10, 1, 0.1 etcétera.

Observe que para definir cada hoja se emplea un solo dígito y que para construir el diagrama sólo se usaron los primeros tres dígitos de cada dato. En la parte superior del diagrama se ha especificado que la Unidad de hoja = 10. Para ilustrar cómo se interpretan los datos de este diagrama considere el primer tallo 15 y su hoja correspondiente 6. Al unir estos números obtiene 156. Para lograr una aproximación al dato original es necesario multiplicar este número por 10, el valor de la *unidad de hoja*. Por tanto, $156 \times 10 = 1560$ es una aproximación al dato original empleado para construir el diagrama de tallo y hoja. Aunque a partir de este diagrama no es posible reconstruir los datos exactos, la convención de usar un solo dígito para cada hoja, permite construir diagramas de tallo y hojas con datos que tengan un gran número de dígitos. En diagramas de tallo y hojas en los que no se especifica la unidad de hoja, se supone que la unidad es 1.

Ejercicios

Métodos

22. Con los datos siguientes construya un diagrama de tallo y hojas.

70	72	75	64	58	83	80	82
76	75	68	65	57	78	85	72

Autoexamen

23. Con los datos siguientes construya un diagrama de tallo y hojas.

11.3	9.6	10.4	7.5	8.3	10.5	10.0
9.3	8.1	7.7	7.5	8.4	6.3	8.8

24. Con los datos siguientes construya un diagrama de tallo y hojas. Use 10 como unidad de hoja.

1161	1206	1478	1300	1604	1725	1361	1422
1221	1378	1623	1426	1557	1730	1706	1689

Aplicaciones



25. Un psicólogo elabora una nueva prueba de inteligencia para adultos. Aplica la prueba a 20 individuos y obtiene los datos siguientes.

114	99	131	124	117	102	106	127	119	115
98	104	144	151	132	106	125	122	118	118

Construya un diagrama de tallo y hojas.

26. La asociación estadounidense de inversionistas individuales realiza una investigación anual sobre intermediarios de descuento. Las siguientes son las comisiones en una muestra de 24 intermediarios (*AAII Journal*, enero de 2003). Estas son dos tipos de operaciones con asistencia de 100 acciones a $50 cada una y una operación en línea de 500 acciones a $50 cada una.



Corredor	Operación con asistencia de 100 acciones $50/ acción	Operación en línea de 500 acciones a $50 /acción	Corredor	Operación con asistencia de 100 acciones $50/ acción	Operación en línea de 500 acciones a $50/acción
Accutrade	30.00	29.95	Merrill Lynch Direct	50.00	29.95
Ameritrade	24.99	10.99	Muriel Siebert	45.00	14.95
Banc of America	54.00	24.95	NetVest	24.00	14.00
Brown & Co.	17.00	5.00	Recom Securities	35.00	12.95
Charles Schwab	55.00	29.95	Scottrade	17.00	7.00
CyberTrader	12.95	9.95	Sloan Securities	39.95	19.95
E*TRADE Securities	49.95	14.95	Strong Investments	55.00	24.95
First Discount	35.00	19.75	TD Waterhouse	45.00	17.95
Freedom Investments	25.00	15.00	T. Rowe Price	50.00	19.95
Harrisdirect	40.00	20.00	Vanguard	48.00	20.00
Investors National	39.00	62.50	Wall Street Discount	29.95	19.95
MB Trading	9.95	10.55	York Securities	40.00	36.00

a. Redondee los precios al dólar más cercano y elabore un diagrama de tallo y hojas de las 100 acciones a $50 por acción. Haga un comentario sobre la información que obtuvo acerca de estos precios.
b. Redondee los precios al dólar más cercano y elabore un diagrama de tallo y hojas de las 500 acciones a $50 por acción. Haga un comentario sobre estos precios.

27. La mayor parte de los centros turísticos importantes de esquí de Estados Unidos ofrecen programas familiares con clases de esquí para niños. Por lo general proporcionan 4 a 6 horas de clase con un instructor certificado. A continuación se presentan las cuotas diarias en 15 centros turísticos. (*The Wall Street Journal,* 20 de enero de 2006).

Centro turístico	Ubicación	Cuota diaria	Centro turístico	Ubicación	Cuota diaria
Beaver Creek	Colorado	$ 137	Okemo	Vermont	$ 86
Deer Valley	Utah	115	Park City	Utah	145
Diamond Peak	California	95	Butternut	Massachusetts	75
Heavenly	California	145	Steamboat	Colorado	98
Hunter	New York	79	Stowe	Vermont	104
Mammoth	California	111	Sugar Bowl	California	100
Mount Sunapee	New Hampshire	96	Whistler-Blackcomb	British Columbia	104
Mount Bachelor	Oregon	83			

a. Con estos datos elabore un diagrama de tallo y hojas.
b. Interprete el diagrama de tallo y hojas en términos de lo que expresa de las cuotas diarias de estos programas.

28. Para un maratón (13.1 millas) en Florida en 2004 hubo 1228 registrados (*Naples Daily News,* 17 de enero de 2004). Para esta competencia hubo seis grupos de edades. Los datos siguientes son las edades encontradas en una muestra de 40 participantes.



Marathon

49	33	40	37	56
44	46	57	55	32
50	52	43	64	40
46	24	30	37	43
31	43	50	36	61
27	44	35	31	43
52	43	66	31	50
72	26	59	21	47

a. Realice un diagrama expandido de tallo y hojas.
b. ¿En qué grupo de edad hubo más participantes?
c. ¿Qué edad se presenta con más frecuencia?
d. En un artículo del *Naples Daily News* se hace énfasis sobre la cantidad de corredores de veintitantos años. ¿Qué porcentaje de los corredores pertenecían al grupo de veintitantos años? ¿Cuál supone qué era el tema del artículo?



Tabulaciones cruzadas y diagramas de dispersión

Las tabulaciones cruzadas y los diagramas de dispersión son empleados para presentar un resumen de datos, de tal manera que revele la relación entre las dos variables.

Este capítulo, hasta ahora, se ha concentrado en los métodos tabulares y gráficos empleados para resumir datos de una *sola variable.* Con frecuencia, los directivos o quienes deben tomar decisiones requieren métodos tabulares o gráficos que les ayuden a entender la *relación entre dos variables*. La tabulación cruzada y los diagramas de dispersión son dos métodos de este tipo.

Tabulación cruzada

Una **tabulación cruzada** es un resumen tabular de los datos de dos variables. El uso de la tabulación cruzada se ilustrará con los datos de la aplicación siguiente, que se basan en datos de *Zagat's Restaurant Review*. Se recolectaron los datos correspondientes a la calidad y precios de 300 restaurantes en el área de Los Ángeles. La tabla 2.9 muestra los datos de los 10 primeros restaurantes. Se presentan los datos de calidad y precio característicos de estos restaurantes. La calidad es una variable cualitativa que tiene como categorías bueno, muy bueno y excelente. El precio es una variable cuantitativa que va desde $10 hasta $49.

En la tabla 2.10 se muestra una tabulación cruzada con los datos de esta aplicación. El encabezado de la primera columna y el primer renglón definen las clases para las dos variables. Los encabezados de los renglones en el margen izquierdo (buena, muy buena y excelente) corresponden a las tres categorías de calidad. Los encabezados de las columnas ($10–19, $20–29, $30–39 y

TABLA 2.9 EVALUACIÓN DE LA CALIDAD Y PRECIOS DE 300 RESTAURANTES DE LOS ÁNGELES



Restaurant

Restaurante	Calidad	Precio
1	Bueno	18
2	Muy bueno	22
3	Bueno	28
4	Excelente	38
5	Muy bueno	33
6	Bueno	28
7	Muy bueno	19
8	Muy bueno	11
9	Muy bueno	23
10	Bueno	13
.	.	.
.	.	.
.	.	.

$40–49) corresponden a las cuatro clases de la variable precio. Para cada restaurante de la muestra se tiene el nivel de calidad y el precio. Por tanto, a cada restaurante de la muestra le corresponde una celda en un renglón y en una columna de la tabla. Por ejemplo, si el restaurante 5 tiene muy buena calidad y su precio es $33, a este restaurante le corresponde el renglón 2 y la columna 3 de la tabla 2.10. Así que para elaborar una tabulación cruzada, simplemente se cuenta el número de restaurantes que pertenecen a cada una de las celdas de la tabla de tabulación cruzada.

La tabla 2.10 muestra que la mayor parte de los restaurantes de la muestra (64) tienen muy buena calidad y su precio está en el intervalo $20–29. También se ve que sólo dos restaurantes tienen una calidad excelente y un precio en el intervalo $10–19. Así es posible hacer interpretaciones semejantes con el resto de las frecuencias. Observe además que en el margen derecho y en el renglón inferior de la tabulación cruzada aparecen las distribuciones de frecuencia de la calidad y de los precios, por separado. En la distribución de frecuencia de la calidad, en el margen derecho, se observa que hay 84 restaurantes buenos, 150 muy buenos y 66 restaurantes excelentes. De manera semejante, en el renglón inferior se tiene la distribución de frecuencia de la variable precios.

Al dividir los totales del margen derecho de la tabulación cruzada entre el total de esa columna se obtienen distribuciones de frecuencia relativa y frecuencia porcentual de la variable calidad.

Calidad	Frecuencia relativa	Frecuencia porcentual
Bueno	0.28	28
Muy bueno	0.50	50
Excelente	0.22	22
Total	1.00	100

TABLA 2.10 TABULACIÓN CRUZADA DE CALIDAD Y PRECIO DE 300 RESTAURANTES DE LOS ÁNGELES

	Precio				
Calidad	**$10–19**	**$20–29**	**$30–39**	**$40–49**	**Total**
Buena	42	40	2	0	84
Muy buena	34	64	46	6	150
Excelente	2	14	28	22	66
Total	78	118	76	28	300

En esta distribución de frecuencia porcentual se observa que 28% de los restaurantes son calificados como buenos, 50% como muy buenos y 22% excelentes.

Si divide los totales del renglón inferior de la tabulación cruzada entre el total de ese renglón obtiene distribuciones de frecuencia relativa y de frecuencia porcentual de los precios.

Precio	Frecuencia relativa	Frecuencia porcentual
$10–19	0.26	26
$20–29	0.39	39
$30–39	0.25	25
$40–49	0.09	9
Total	1.00	100

Observe que la suma de los valores en cada columna no tiene correspondencia exacta con el total de la columna debido a que los valores que se suman han sido redondeados. En esta distribución de frecuencia porcentual 26% de los precios se encuentran en la clase de los precios más bajos, 39% se encuentran en la clase siguiente, etcétera.

Las distribuciones de frecuencia y de frecuencia relativa obtenidas de los márgenes de las tabulaciones cruzadas proporcionan información de cada una de las variables por separado, pero no dan ninguna luz acerca de la relación entre las variables. El principal valor de una tabulación cruzada es que permite ver la relación entre las variables. Una observación de la tabulación cruzada de la tabla 2.10 es que los precios más altos están relacionados con la mejor calidad de los restaurantes y los precios bajos están relacionados con menor calidad.

Si se convierten las cantidades de una tabulación cruzada en porcentajes de columna o de renglón, se obtiene más claridad sobre la relación entre las variables. En la tabla 2.11 se presentan los porcentajes de renglón, que son el resultado de dividir cada frecuencia de la tabla 2.10 entre el total del renglón correspondiente. Entonces, cada renglón de la tabla 2.11 es una distribución de frecuencia porcentual de los precios en esa categoría de calidad. Entre los restaurantes de menor calidad (buenos), el mayor porcentaje corresponde a los menos caros (50% tiene precios en el intervalo $10–19 y 47.6% en el intervalo $20–29). De los restaurantes de mayor calidad (excelentes), los porcentajes mayores corresponden a los más caros (42.4% tiene precios de $30–39 y 33.4% de $40–49). Así que un precio más elevado está relacionado con una mejor calidad de los restaurantes.

La tabulación cruzada se utiliza mucho para examinar la relación entre dos variables. En la práctica, los informes finales de muchos estudios estadísticos contienen una gran cantidad de tabulaciones cruzadas. En este estudio sobre los restaurantes de Los Ángeles, en la tabulación cruzada se emplea una variable cualitativa (las calidades) y una cuantitativa (los precios). También se elaboran tabulaciones cruzadas con dos variables cualitativas o cuantitativas. Cuando se usan variables cuantitativas, primero es necesario crear las clases para los valores de las variables. Por ejemplo, en el caso de los restaurantes se agruparon los precios en cuatro categorías ($10–19, $20–29, $30–39 y $40–49).

TABLA 2.11 PORCENTAJES DE RENGLÓN DE CADA CATEGORÍA DE CALIDAD

	Precio				
Calidad	**$10–19**	**$20–29**	**$30–39**	**$40–49**	**Total**
Buena	50.0	47.6	2.4	0.0	100
Muy buena	22.7	42.7	30.6	4.0	100
Excelente	3.0	21.2	42.4	33.4	100

Paradoja de Simpson

Es posible combinar o agregar los datos de dos o más tabulaciones cruzadas para obtener una tabulación cruzada resumida que muestre la relación entre dos variables. En tales casos hay que tener mucho cuidado al sacar conclusiones acerca de la relación entre las dos variables de la tabulación cruzada agregada. En algunos casos las conclusiones obtenidas de la tabulación cruzada agregada se invierten por completo al observar los datos no agregados, situación conocida como **paradoja de Simpson**. Para ilustrar la paradoja de Simpson, se proporciona un ejemplo en el que se analizan las sentencias de dos jueces en dos tipos de tribunales.

Los jueces Ron Luckett y Dennis Kendall, presidieron los tres últimos años dos tipos de tribunales, de primera instancia y municipal. Algunas de las sentencias por ellos dictadas fueron apeladas. En la mayor parte de los casos los tribunales de apelación ratificaron las sentencias, pero en algunos casos fueron revocadas. Para cada juez se elabora una tabulación cruzada con las variables: sentencia (ratificada o revocada) y tipo de tribunal (de primera instancia y municipal). Suponga que después se combinan las dos tabulaciones cruzadas agregando los datos de los dos tipos de tribunales. La tabulación cruzada agregada que se obtiene tiene dos variables: sentencia (ratificada o revocada) y juez (Luckett o Kendall). En esta tabulación cruzada para cada uno de los jueces se da la cantidad de sentencias que fueron ratificadas y la cantidad de sentencias que fueron revocadas. En la tabla siguiente se presentan estos resultados junto a los porcentajes de columna entre paréntesis al lado de cada valor.

	Juez		
Sentencia	**Luckett**	**Kendall**	**Total**
Ratificada	129 (86%)	110 (88%)	239
Revocada	21 (14%)	15 (12%)	36
Total (%)	150 (100%)	125 (100%)	275

Al analizar la columna de porcentajes resulta que 14% de las sentencias del juez Luckett fueron revocadas, pero del juez Kendall sólo 12% de las sentencias lo fueron. Por tanto, el juez Kendall tuvo un mejor desempeño, ya que de sus sentencias se ratificó un porcentaje mayor. Sin embargo, de esta conclusión surge un problema.

En la tabla siguiente se muestran los casos atendidos por cada uno de los jueces en los dos tribunales; aquí también se dan los porcentajes entre paréntesis al lado de los valores.

Juez Luckett

Sentencia	Tribunal de primera instancia	Tribunal municipal	Total
Ratificada	29 (91%)	100 (85%)	129
Revocada	3 (9%)	18 (15%)	21
Total (%)	32 (100%)	118 (100%)	150

Juez Kendall

Sentencia	Tribunal de primera instancia	Tribunal municipal	Total
Ratificada	90 (90%)	20 (80%)	110
Revocada	10 (10%)	5 (20%)	15
Total (%)	100 (100%)	25 (100%)	125

Respecto de los porcentajes de Luckett, en el tribunal de primera instancia 91% de sus sentencias fueron ratificadas y en el tribunal municipal 85% lo fueron. En cuanto a los porcentajes de Kendall, 90% de sus sentencias del tribunal de primera instancia y 80% del tribunal municipal fueron ratificadas. Al comparar los porcentajes de columna de los dos jueces, es obvio que el juez Luckett tuvo un mejor desempeño en ambos tribunales que el Juez Kendall. Esto contradice las conclusiones obtenidas al agregar los datos de los dos tribunales en la primera tabulación cruzada. Se pensó que el juez Kendall tenía un mejor desempeño. Este ejemplo ilustra la paradoja de Simpson.

La primera tabulación cruzada se obtuvo agregando los datos de los dos tribunales de dos tabulaciones cruzadas. Observe que los dos jueces tuvieron porcentajes mayores de sentencias revocadas en las sentencias del tribunal municipal que en las del tribunal de primera instancia. Como el juez Luckett tuvo un porcentaje mayor de casos del tribunal municipal, los datos agregados favorecieron al juez Kendall. Sin embargo, si presta atención a las tabulaciones cruzadas de cada uno de los jueces, es claro que el juez Luckett tuvo un mejor desempeño. Por tanto, en la primera tabulación cruzada el *tipo de tribunal* es una variable oculta que no debe ser ignorada al evaluar el desempeño de estos dos jueces.

Debido a la paradoja de Simpson, es necesario tener mucho cuidado al sacar conclusiones cuando se usan datos agregados. Antes de cualquier conclusión acerca de la relación entre dos variables, en una tabulación cruzada en la que se usan datos agregados, es preciso investigar si no existen variables ocultas que afecten los resultados.

Diagrama de dispersión y línea de tendencia

Un **diagrama de dispersión** es una representación gráfica de la relación entre dos variables cuantitativas y una **línea de tendencia** es una línea que da una aproximación de la relación. Como ejemplo, considere la relación publicidad/ventas en una tienda de equipos de sonido. Durante los últimos tres meses, en 10 ocasiones la tienda apareció en comerciales de televisión, en el fin de semana, para promover sus ventas. Los directivos quieren investigar si hay relación entre el número de comerciales emitidos el fin de semana y las ventas en la semana siguiente. En la tabla 2.12 se presentan datos muestrales de las 10 semanas dando las ventas en cientos de dólares.

En la figura 2.7 aparece el diagrama de dispersión y la línea de tendencia* de los datos de la tabla 2.12. El número de comerciales (x) aparece en el eje horizontal y las ventas (y) en el eje vertical. En la semana 1, $x = 2$ y $y = 50$. En el diagrama de dispersión se grafica un punto con estas coordenadas. Para las otras nueve semanas se grafican puntos similares. Observe que en dos semanas sólo hubo un comercial, en otras dos semanas hubo dos comerciales, etcétera.

De nuevo, respecto de la figura 2.7, se observa una relación positiva entre el número de comerciales y las ventas. Más ventas corresponden a más comerciales. La relación no es perfecta ya que los puntos no trazan una línea recta. Sin embargo, el patrón que siguen los puntos y la línea de tendencia indican que la relación es positiva.

En la figura 2.8 se muestran los patrones de los diagramas de dispersión y el tipo de relación que sugieren. La gráfica arriba a la izquierda representa una relación positiva parecida a la del

TABLA 2.12 DATOS MUESTRALES DE UNA TIENDA DE EQUIPOS DE SONIDO



Stereo

Semana	Número de comerciales x	Ventas (en cientos de dólares) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

*La ecuación de la línea de tendencia es $y = 36.15 + 4.95x$. La pendiente de la línea de tendencia es 4.95 y la intersección con el eje y (el punto en que la recta interseca el eje y) es 36.15. La interpretación de la pendiente y de la intersección con el eje y de una línea de tendencia lineal lo verá con detalle en el capítulo 12, cuando estudie la regresión lineal simple.

FIGURA 2.7 DIAGRAMA DE DISPERSIÓN Y LÍNEA DE TENDENCIA DE LA TIENDA DE EQUIPOS DE SONIDO

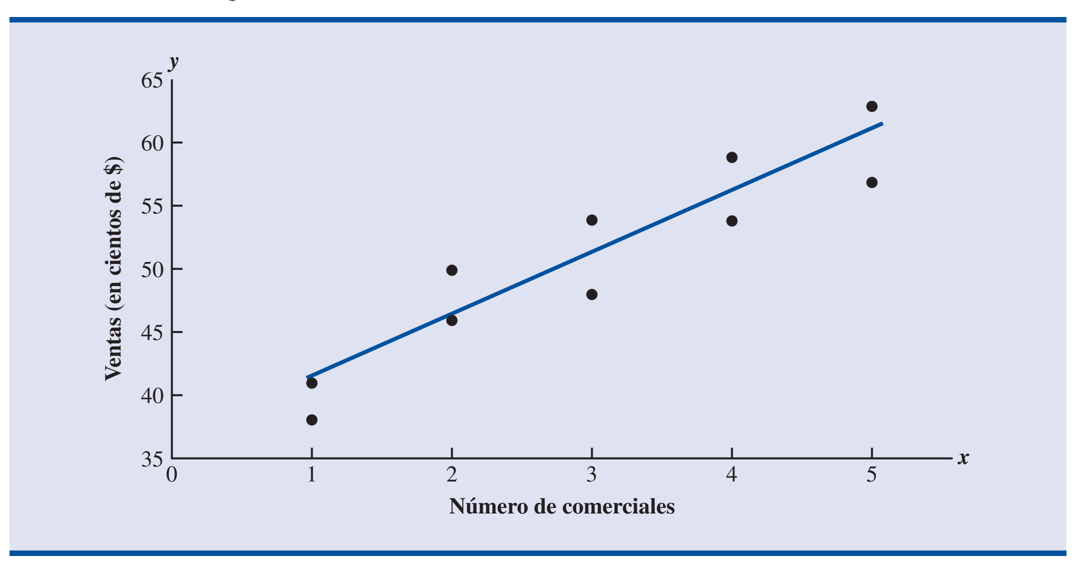
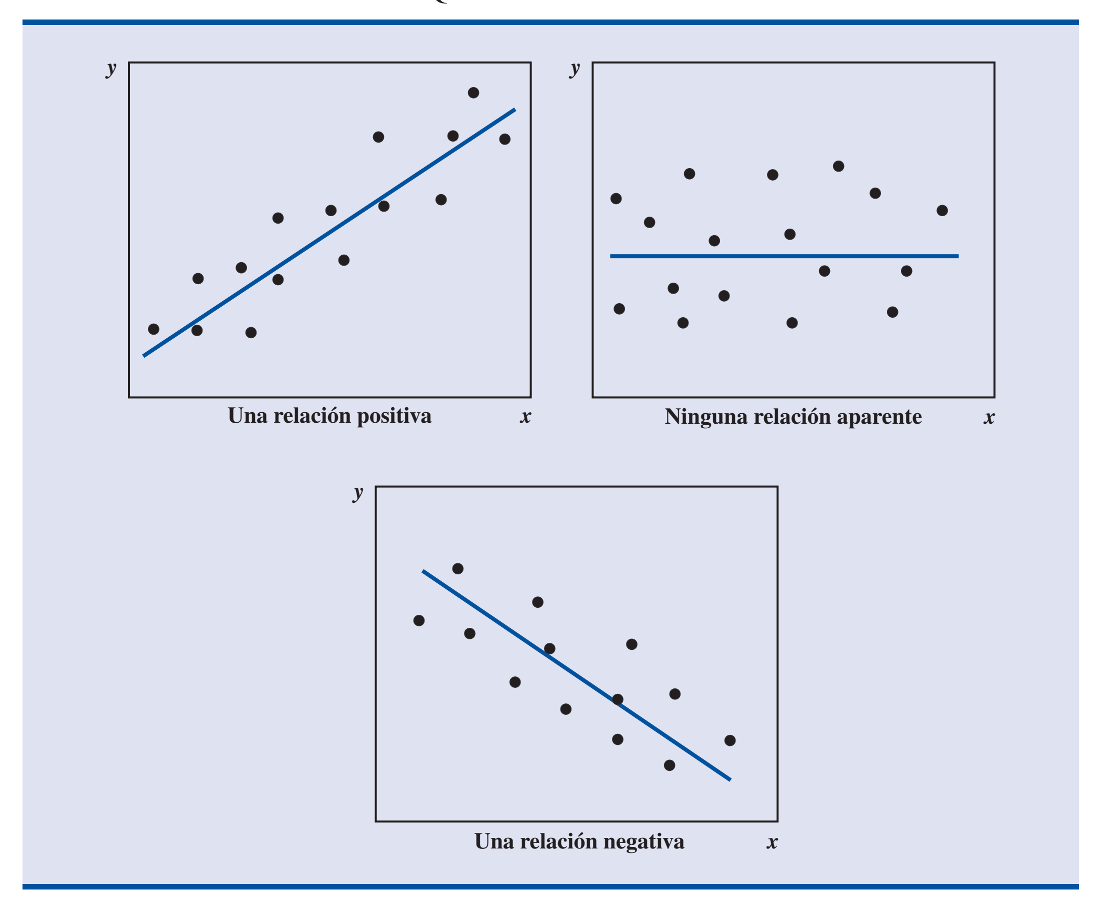


FIGURA 2.8 TIPOS DE RELACIÓN QUE APARECEN EN LOS DIAGRAMAS DE DISPERSIÓN



ejemplo de la cantidad de comerciales y las ventas. En la gráfica de arriba a la derecha no aparece ninguna relación entre las dos variables. La gráfica inferior representa una relación negativa en la que *y* tiende a disminuir a medida que *x* aumenta.

Ejercicios

Métodos



29. Los siguientes son datos de 30 observaciones en las que intervienen dos variables, *x* y *y*. Las categorías para *x* son A, B, y C; para *y* son 1 y 2.



Crosstab

Observación	*x*	*y*	Observación	*x*	*y*
1	A	1	16	B	2
2	B	1	17	C	1
3	B	1	18	B	1
4	C	2	19	C	1
5	B	1	20	B	1
6	C	2	21	C	2
7	B	1	22	B	1
8	C	2	23	C	2
9	A	1	24	A	1
10	B	1	25	B	1
11	A	1	26	C	2
12	B	1	27	C	2
13	C	2	28	A	1
14	C	2	29	B	1
15	C	2	30	B	2

a. Con estos datos elabore una tabulación cruzada en la que *x* sea la variable para los renglones y *y* para las columnas.
b. Calcule los porcentajes de los renglones.
c. Calcule los porcentajes de las columnas.
d. ¿Cuál es la relación, si hay alguna, entre las variables *x* y *y*?





30. Las siguientes 20 observaciones corresponden a 20 variables cuantitativas, *x* y *y*.



Scatter

Observación	*x*	*y*	Observación	*x*	*y*
1	−22	22	11	−37	48
2	−33	49	12	34	−29
3	2	8	13	9	−18
4	29	−16	14	−33	31
5	−13	10	15	20	−16
6	21	−28	16	−3	14
7	−13	27	17	−15	18
8	−23	35	18	12	17
9	14	−5	19	−20	−11
10	3	−3	20	−7	−22

a. Elabore un diagrama de dispersión para la relación entre *x* y *y*.
b. ¿Cuál es la relación, si hay alguna, entre *x* y *y*?

Aplicaciones

31. En la siguiente tabulación cruzada se muestra el ingreso familiar de acuerdo con el nivel de estudios del jefe de familia, (*Statistical Abstract of the United States, 2002*).

	Ingreso por familia (en miles de dólares)					
Nivel de estudios	**Menos de 25**	**25.0–49.9**	**50.0–74.9**	**75.0–99.9**	**100 o más**	**Total**
No terminó secundaria	9 285	4 093	1 589	541	354	15 862
Terminó secundaria	10 150	9 821	6 050	2 737	2 028	30 786
Parte de bachillerato	6 011	8 221	5 813	3 215	3 120	26 380
Título universitario	2 138	3 985	3 952	2 698	4 748	17 521
Posgrado	813	1 497	1 815	1 589	3 765	9 479
Total	28 397	27 617	19 219	10 780	14 015	100 028

a. Calcule los porcentajes por renglón e identifique las distribuciones de frecuencia porcentual del ingreso en los hogares en que el jefe de familia terminó secundaria y en los hogares en que el jefe de familia tiene un título universitario.
b. ¿Qué porcentaje de las familias en que el jefe de familia terminó secundaria gana $75 000 o más? ¿Qué porcentaje de las familias en que el jefe de familia tienen un título universitario gana 75 000 o más?
c. Con los ingresos de los hogares en que el jefe de familia terminó secundaria elabore un histograma de la frecuencia porcentual, y otro con los ingresos de las familias en que el jefe de familia tiene un grado universitario. ¿Se observa alguna relación clara entre el ingreso familiar y el nivel de educación?

32. Consulte la tabulación cruzada del ingreso familiar de acuerdo con el nivel de estudios del ejercicio 31.
a. Calcule los porcentajes e identifique las distribuciones de frecuencia porcentual. ¿Qué porcentaje de jefes de familia no terminó la secundaria?
b. ¿Qué porcentaje de los hogares que perciben $100 000 o más tienen como jefe de familia a una persona con un posgrado? ¿ Qué porcentaje de los hogares que tienen como jefe de familia a una persona con un posgrado perciben más de $100 000? ¿Por qué son diferentes estos dos porcentajes?
c. Compare las distribuciones de frecuencia porcentual de aquellos hogares que perciben "Menos que 25", "100 o más" y del "Total". Haga un comentario sobre la relación entre ingreso familiar y nivel de estudios del jefe de familia.

33. Hace poco los administradores de un campo de golf recibieron algunas quejas acerca de las condiciones de los *greens*. Varios jugadores se quejaron de que estaban demasiado rápidos. En lugar de reaccionar a los comentarios de unos cuantos, la asociación de golf realizó un sondeo con 100 jugadoras y 100 jugadores. Los resultados del sondeo se presentan a continuación.

Jugadores

	Condición de los *greens*	
Hándicap	**Demasiado rápido**	**Bien**
Menos de 15	10	40
15 o más	25	25

Jugadoras

	Condición de los greens	
Hándicap	**Demasiado rápido**	**Bien**
Menos de 15	1	9
15 o más	39	51

a. Combine estas dos tabulaciones cruzadas utilizando como encabezados de renglón Jugadores y Jugadoras y como encabezados de columnas Demasiado rápido y Bien. ¿En qué grupo se encuentra el mayor porcentaje de los que dicen que los *greens* están demasiado rápidos?

b. Vuelva a las tabulaciones cruzadas iniciales. De los jugadores con bajo hándicap (mejores jugadores), ¿en qué grupo (jugadoras o jugadores) se encuentra un porcentaje mayor de quienes dicen que los *greens* están demasiado rápidos?

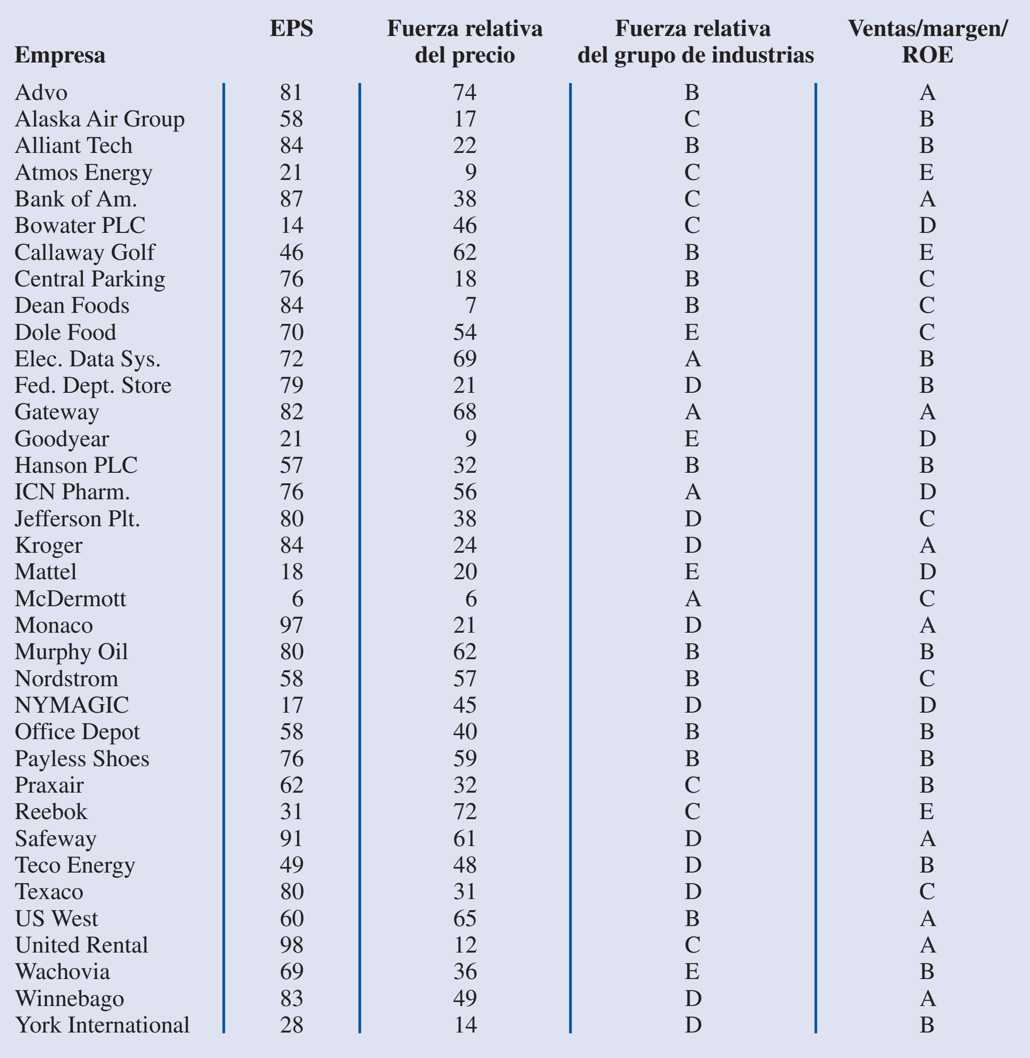
c. Regrese a las tabulaciones cruzadas iniciales. De los jugadores con alto hándicap, ¿en qué grupo (jugadoras o jugadores) se encuentra un porcentaje mayor para quienes los *greens* están demasiado rápidos?

d. ¿Qué conclusiones obtiene acerca de mujeres y hombres respecto a la velocidad de los *greens*? ¿Las conclusiones que obtuvo en el inciso a son consistentes con los incisos b y c? Explique cualquier inconsistencia aparente.

34. En la tabla 2.13 se presentan datos financieros de 36 empresas de una muestra cuyas acciones cotizan en la bolsa de valores de Nueva York (*Investor's Business Daily,* 7 de abril de 2000). Los datos de la columna Ventas/margen/ROE son evaluaciones financieras compuestas que se basan en la tasa de crecimiento de las ventas de una empresa, su margen de ganancia y su rendimiento de los activos (ROE *return on capital employed*). La calificación EPS es una medida del crecimiento por acción.



TABLA 2.13 DATOS FINANCIEROS DE 36 EMPRESAS QUE CONFORMAN UNA MUESTRA



Empresa	EPS	Fuerza relativa del precio	Fuerza relativa del grupo de industrias	Ventas/margen/ROE
Advo	81	74	B	A
Alaska Air Group	58	17	C	B
Alliant Tech	84	22	B	B
Atmos Energy	21	9	C	E
Bank of Am.	87	38	C	A
Bowater PLC	14	46	C	D
Callaway Golf	46	62	B	E
Central Parking	76	18	B	C
Dean Foods	84	7	B	C
Dole Food	70	54	E	C
Elec. Data Sys.	72	69	A	B
Fed. Dept. Store	79	21	D	B
Gateway	82	68	A	A
Goodyear	21	9	E	D
Hanson PLC	57	32	B	B
ICN Pharm.	76	56	A	D
Jefferson Plt.	80	38	D	C
Kroger	84	24	D	A
Mattel	18	20	E	D
McDermott	6	6	A	C
Monaco	97	21	D	A
Murphy Oil	80	62	B	B
Nordstrom	58	57	B	C
NYMAGIC	17	45	D	D
Office Depot	58	40	B	B
Payless Shoes	76	59	B	B
Praxair	62	32	C	B
Reebok	31	72	C	E
Safeway	91	61	D	A
Teco Energy	49	48	D	B
Texaco	80	31	D	C
US West	60	65	B	A
United Rental	98	12	C	A
Wachovia	69	36	E	B
Winnebago	83	49	D	A
York International	28	14	D	B



Fuente: Investor's Business Daily, 7 de abril de 2000.

a. Elabore una tabulación cruzada con los datos Ventas/margen/ROE (renglones) y EPS (columnas). Para el EPS emplee las clases 0–19, 20–39, 40–59, 60–79 y 80–99.
b. Calcule los porcentajes de las columnas y haga un comentario sobre la relación entre las variables.

35. Regrese a la tabla 2.13.
a. Elabore una tabulación cruzada con los datos Ventas/margen/ROE y Fuerza relativa del grupo de industrias.
b. Elabore una distribución de frecuencia de los datos Ventas/margen/ROE.
c. Elabore una distribución de frecuencia de los datos Fuerza relativa del grupo de industrias.
d. ¿Le ayudó la tabulación cruzada en la elaboración de las distribuciones de frecuencia de los incisos b y c?

36. De nuevo, a la tabla 2.13.
a. Elabore un diagrama de dispersión con los datos EPS y Fuerza relativa del precio.
b. Haga un comentario sobre la relación entre las variables. (El significado del EPS se describe en el ejercicio 34. La Fuerza relativa del precio es una medida de la variación en el precio de una acción en los últimos 12 meses. Valores altos indican gran variación.)

37. La National Football League de Estados Unidos evalúa a los candidatos posición por posición con una escala que va de 5 a 9. La evaluación se interpreta como sigue: 8–9 debe empezar el primer año; 7.0–7.9 debe empezar; 6.0–6.9 será un apoyo para el equipo, y 5.0–5.9 puede pertenecer al club y contribuir. En la tabla 2.14 se presentan posición, peso, tiempo (segundos en correr 40 yardas), y evaluación de 40 candidatos (*USA Today,* 14 de abril de 2000).
a. Con los datos posición (renglones) y tiempo (columnas) elabore una tabulación cruzada. Para el tiempo emplee las clases 4.00–4.49, 4.50–4.99, 5.00–5.49 y 5.50–5.99.
b. Haga un comentario acerca de la relación entre posición y tiempo, con base en la tabulación cruzada que elaboró en el inciso a.
c. Con los datos tiempo y calificación obtenida en la evaluación elabore un diagrama de dispersión, coloque la calificación obtenida en la evaluación en el eje vertical.
d. Haga un comentario sobre la relación entre tiempo y calificación obtenida en la evaluación.

Resumen

Un conjunto de datos, aunque sea de tamaño modesto, es difícil de interpretar con los datos tal y como se han recolectado. Los métodos tabulares y los métodos gráficos permiten organizar y resumir los datos para que muestren algún patrón y sean factibles de interpretación. Para resumir datos cualitativos se presentaron las distribuciones de frecuencia, de frecuencia relativa y las de frecuencia porcentual, las gráficas de barras y las gráficas de pastel. Las distribuciones de frecuencia, de frecuencia relativa, de frecuencia porcentual, los histogramas, las distribuciones de frecuencia acumulada, de frecuencia relativa acumulada, de frecuencia porcentual acumulada y las ojivas se presentaron como métodos para resumir datos cuantitativos. Los diagramas de tallo y hojas son una técnica para el análisis exploratorio de datos que se usa para resumir datos cuantitativos. La tabulación cruzada se presentó como un método para resumir datos para dos variables. Los diagramas de dispersión se presentaron como un método gráfico para mostrar la relación entre dos variables cuantitativas. En la figura 2.9 se resumen los métodos tabulares y gráficos que se presentaron en este capítulo.

Cuando se tienen grandes conjuntos de datos es indispensable usar paquetes de software para la elaboración de resúmenes tabulares o gráficos de los datos. En los dos apéndices de este capítulo se explica el uso de Minitab y de Excel con tal propósito.



NFL

TABLA 2.14 DATOS DE 40 CANDIDATOS A LA NATIONAL FOOTBALL LEAGUE DE ESTADOS UNIDOS

Observación	Nombre	Posición	Peso	Tiempo	Evaluación
1	Peter Warrick	Receptor abierto	194	4.53	9
2	Plaxico Burress	Receptor abierto	231	4.52	8.8
3	Sylvester Morris	Receptor abierto	216	4.59	8.3
4	Travis Taylor	Receptor abierto	199	4.36	8.1
5	Laveranues Coles	Receptor abierto	192	4.29	8
6	Dez White	Receptor abierto	218	4.49	7.9
7	Jerry Porter	Receptor abierto	221	4.55	7.4
8	Ron Dugans	Receptor abierto	206	4.47	7.1
9	Todd Pinkston	Receptor abierto	169	4.37	7
10	Dennis Northcutt	Receptor abierto	175	4.43	7
11	Anthony Lucas	Receptor abierto	194	4.51	6.9
12	Darrell Jackson	Receptor abierto	197	4.56	6.6
13	Danny Farmer	Receptor abierto	217	4.6	6.5
14	Sherrod Gideon	Receptor abierto	173	4.57	6.4
15	Trevor Gaylor	Receptor abierto	199	4.57	6.2
16	Cosey Coleman	Guardia	322	5.38	7.4
17	Travis Claridge	Guardia	303	5.18	7
18	Kaulana Noa	Guardia	317	5.34	6.8
19	Leander Jordan	Guardia	330	5.46	6.7
20	Chad Clifton	Guardia	334	5.18	6.3
21	Manula Savea	Guardia	308	5.32	6.1
22	Ryan Johanningmeir	Guardia	310	5.28	6
23	Mark Tauscher	Guardia	318	5.37	6
24	Blaine Saipaia	Guardia	321	5.25	6
25	Richard Mercier	Guardia	295	5.34	5.8
26	Damion McIntosh	Guardia	328	5.31	5.3
27	Jeno James	Guardia	320	5.64	5
28	Al Jackson	Guardia	304	5.2	5
29	Chris Samuels	Tacle ofensivo	325	4.95	8.5
30	Stockar McDougle	Tacle ofensivo	361	5.5	8
31	Chris McIngosh	Tacle ofensivo	315	5.39	7.8
32	Adrian Klemm	Tacle ofensivo	307	4.98	7.6
33	Todd Wade	Tacle ofensivo	326	5.2	7.3
34	Marvel Smith	Tacle ofensivo	320	5.36	7.1
35	Michael Thompson	Tacle ofensivo	287	5.05	6.8
36	Bobby Williams	Tacle ofensivo	332	5.26	6.8
37	Darnell Alford	Tacle ofensivo	334	5.55	6.4
38	Terrance Beadles	Tacle ofensivo	312	5.15	6.3
39	Tutan Reyes	Tacle ofensivo	299	5.35	6.1
40	Greg Robinson-Ran	Tacle ofensivo	333	5.59	6

FIGURA 2.9 MÉTODOS TABULARES Y GRÁFICOS PARA RESUMIR DATOS

Glosario

Datos cualitativos Etiquetas o nombres que se usan para identificar las categorías de elementos semejantes.

Datos cuantitativos Valores numéricos que indican cuánto o cuántos.

Distribución de frecuencia Resumen tabular de datos que muestra el número (frecuencia) de los datos que pertenecen a cada una de varias clases disyuntas.

Distribución de frecuencia relativa Resumen tabular de datos que muestra la proporción o la fracción de datos propios de cada una de varias clases disyuntas.

Distribución de frecuencia porcentual Resumen tabular de datos que muestra el porcentaje de datos que corresponden a cada una de varias clases disyuntas.

Gráfica de barras Gráfica para representar datos cualitativos que hayan sido resumidos en una distribución de frecuencia, de frecuencia relativa o de frecuencia porcentual.

Gráfica de pastel Gráfica para representar datos resumidos mediante una distribución de frecuencia relativa y que se basa en la subdivisión de un círculo en sectores que corresponden a la frecuencia relativa de las clases.

Punto medio de clase Valor que se encuentra a la mitad entre el límite de clase inferior y el límite de clase superior.

Gráfica de puntos Gráfica que resume datos mediante la cantidad de puntos sobre los valores de los datos que se encuentran en un eje horizontal.

Histograma Representación gráfica de una distribución de frecuencia, de frecuencia relativa o de frecuencia porcentual que se construye colocando los intervalos de clase sobre un eje horizontal y la frecuencia, la frecuencia relativa o la frecuencia porcentual sobre un eje vertical.

Distribución de frecuencia acumulada Síntesis tabular de datos cuantitativos, en la que se muestra el número de datos que son menores o iguales que el límite superior de cada clase.
Distribución de frecuencia relativa acumulada Resumen tabular de datos cuantitativos, en el que se muestra la proporción o fracción de datos que son menores o iguales que el límite superior de cada clase.
Distribución de frecuencia porcentual acumulada Síntesis tabular de datos cuantitativos, en la que se muestra el porcentaje de datos que son menores o iguales que el límite superior de cada clase.
Ojiva Gráfica de una distribución acumulada.
Análisis exploratorio de datos Métodos en los que se emplean cálculos aritméticos sencillos y gráficas fáciles de elaborar para resumir datos en forma rápida.
Diagrama de tallo y hojas Técnica para el análisis exploratorio de datos que tanto ordena por jerarquía datos cuantitativos como proporciona claridad acerca de la forma de la distribución.
Tabulación cruzada Resumen tabular de datos de dos variables. Las clases de una de las variables se representan como renglones; las clases de la otra variable como columnas.
Paradoja de Simpson Conclusiones que se obtienen de dos o más tabulaciones cruzadas y que se invierten cuando se agregan los datos en una sola tabulación cruzada.
Diagrama de dispersión Representación gráfica de la relación entre dos variables cuantitativas. A una variable se le asigna un eje horizontal y a la otra un eje vertical.
Línea de tendencia Línea que da una aproximación de la relación entre dos variables.

Fórmulas clave

Frecuencia relativa

$$\frac{\text{Frecuencia de la clase}}{n} \qquad \textbf{(2.1)}$$

Ancho aproximado de clase

$$\frac{\text{Dato mayor} - \text{Dato menor}}{\text{Número de clases}} \qquad \textbf{(2.2)}$$

Ejercicios complementarios

38. Los cinco automóviles más vendidos en Estados Unidos durante 2003 fueron la camioneta Chevrolet Silverado/C/K, la camioneta Dodge Ram, la camioneta Ford F-Series, el Honda Accord y el Toyota Camry (*Motor Trend,* 2003). En la tabla 2.15 se presenta una muestra de 50 compras de automóviles.



TABLA 2.15 DATOS DE 50 COMPRAS DE AUTOMÓVILES

Silverado	Ram	Accord	Camry	Camry
Silverado	Silverado	Camry	Ram	F-Series
Ram	F-Series	Accord	Ram	Ram
Silverado	F-Series	F-Series	Silverado	Ram
Ram	Ram	Accord	Silverado	Camry
F-Series	Ram	Silverado	Accord	Silverado
Camry	F-Series	F-Series	F-Series	Silverado
F-Series	Silverado	F-Series	F-Series	Ram
Silverado	Silverado	Camry	Camry	F-Series
Silverado	F-Series	F-Series	Accord	Accord

a. Elabore una distribución de frecuencia y otra de frecuencia porcentual.
b. ¿Cuál es la camioneta y el automóvil de pasajeros más vendidos?
c. Haga una gráfica de pastel.

39. El Higher Education Research Institute de UCLA cuenta con estadísticas sobre las áreas que son más elegidas por los estudiantes de nuevo ingreso. Las cinco más elegidas son arte y humanidades (A), administración de negocios (B), ingeniería (E), política (P) y ciencias sociales (S) (*The New York Times Almanac,* 2006). Otras áreas (O), entre las que se encuentran biología, física, ciencias de la computación y educación se agruparon todas en una sola categoría. Las siguientes fueron las áreas elegidas por 64 estudiantes de recién ingreso de una muestra.



S	P	P	O	B	E	O	E	P	O	O	B	O	O	O	A
O	E	E	B	S	O	B	O	A	O	E	O	E	O	B	P
B	A	S	O	E	A	B	O	S	S	O	O	E	B	O	B
A	E	B	E	A	A	P	O	O	E	O	B	B	O	P	B

a. Dé una distribución de frecuencia y otra de frecuencia porcentual.
b. Elabore una gráfica de barras.
c. ¿Que porcentaje de los estudiantes de nuevo ingreso elige una de las cinco áreas más elegidas?
d. ¿Cuál es el área más elegida por los estudiantes de nuevo ingreso? ¿Qué porcentaje de los estudiantes de nuevo ingreso elige esta área?

40. A los 100 mejores entrenadores de golf la revista *Golf Magazine* les preguntó, "¿Cuál es el aspecto más relevante que impide a los jugadores de golf desarrollar todo su potencial?" Las respuestas fueron falta de precisión, técnica de golpe inadecuada, actitud mental inadecuada, falta de energía, práctica insuficiente, tiro al hoyo inadecuado, juego corto inadecuado y estrategia de decisión inadecuada. A continuación se presentan los datos obtenidos (*Golf Magazine*, febrero de 2002):



Actitud mental	Actitud mental	Juego corto	Juego corto	Juego corto
Práctica	Precisión	Actitud mental	Precisión	Tiro al hoyo
Energía	Técnica de golpe	Precisión	Juego corto	Tiro al hoyo
Precisión	Actitud mental	Actitud mental	Precisión	Energía
Precisión	Precisión	Juego corto	Energía	Juego corto
Precisión	Tiro al hoyo	Actitud mental	Estrategia de decisión	Precisión
Juego corto	Energía	Actitud mental	Técnica de golpe	Juego corto
Práctica	Práctica	Actitud mental	Energía	Energía
Actitud mental	Juego corto	Actitud mental	Juego corto	Estrategia de decisión
Precisión	Juego corto	Precisión	Actitud mental	Juego corto
Actitud mental	Tiro al hoyo	Actitud mental	Actitud mental	Tiro al hoyo
Práctica	Tiro al hoyo	Práctica	Juego corto	Tiro al hoyo
Energía	Actitud mental	Juego corto	Práctica	Estrategia de decisión
Precisión	Juego corto	Precisión	Práctica	Tiro al hoyo
Precisión	Juego corto	Precisión	Juego corto	Tiro al hoyo
Precisión	Técnica de golpe	Juego corto	Actitud mental	Práctica
Juego corto	Juego corto	Estrategia de decisión	Juego corto	Juego corto
Práctica	Práctica	Juego corto	Práctica	Estrategia de decisión
Actitud mental	Estrategia de decisión	Estrategia de decisión	Energía	Juego corto
Precisión	Práctica	Práctica	Práctica	Precisión

a. Elabore una distribución de frecuencia y otra de frecuencia porcentual.
b. ¿Cuáles son los aspectos más relevantes que impiden a un jugador de golf desarrollar su potencial?

41. El rendimiento de dividendos son los beneficios anuales que paga una empresa, expresado como porcentaje del precio de una acción (Dividendo/precio de la acción $\times$ 100). En la tabla 2.16 se presenta el rendimiento de dividendos de las empresas del promedio industrial Dow Jones (*The Wall Street Journal*, 3 de marzo de 2006).
a. Haga una distribución de frecuencia y una distribución de frecuencia porcentual.
b. Haga un histograma.
c. Aporte un comentario sobre la forma de la distribución.

TABLA 2.16 RENDIMIENTO DE DIVIDENDOS DE LAS EMPRESAS DEL PROMEDIO INDUSTRIAL DOW JONES.



Empresa	Rendimiento de dividendos	Empresa	Rendimiento de dividendos
AIG	0.9	Home Depot	1.4
Alcoa	2.0	Honeywell	2.2
Altria Group	4.5	IBM	1.0
American Express	0.9	Intel	2.0
AT&T	4.7	Johnson & Johnson	2.3
Boeing	1.6	JPMorgan Chase	3.3
Caterpillar	1.3	McDonald's	1.9
Citigroup	4.3	Merck	4.3
Coca-Cola	3.0	Microsoft	1.3
Disney	1.0	3M	2.5
DuPont	3.6	Pfizer	3.7
ExxonMobil	2.1	Procter & Gamble	1.9
General Electric	3.0	United Technologies	1.5
General Motors	5.2	Verizon	4.8
Hewlett-Packard	0.9	Wal-Mart Stores	1.3

d. ¿Qué indican los resúmenes tabular y gráfico acerca de los dividendos de las empresas del promedio industrial Dow Jones?
e. ¿Qué empresa tiene el más alto rendimiento de dividendos? Si hoy el precio de las acciones de esta empresa es $20 por acción y usted compra 500 acciones, ¿cuál será el ingreso por dividendos que genere anualmente esta inversión?

42. Cada año en Estados Unidos, aproximadamente 1.5 millones de los estudiantes de educación superior presentan un examen de aptitud escolar (SAT, por sus siglas en inglés). Cerca de 80% de las universidades e instituciones de educación superior emplean las puntuaciones obtenidas por los estudiantes en este examen como criterio de admisión (*College Board,* marzo de 2006). A continuación se presentan las puntuaciones obtenidas en las áreas de matemáticas y expresión verbal por una muestra de estudiantes.



1025	1042	1195	880	945
1102	845	1095	936	790
1097	913	1245	1040	998
998	940	1043	1048	1130
1017	1140	1030	1171	1035

a. Presente una distribución de frecuencia y un histograma de estas puntuaciones. La primera clase debe empezar en la puntuación 750 y la amplitud de clase deberá ser 100.
b. Dé un comentario sobre la forma de la distribución.
c. ¿Qué otras observaciones puede hacer acerca de estas puntuaciones con base en los resúmenes tabulares y gráficos?

43. La Asociación estadounidense de inversionistas independientes informa sobre 94 acciones fantasma. El término *fantasma* se refiere a que son acciones de empresas pequeñas o medianas que no son seguidas de cerca por las principales casas de bolsa. A continuación se presenta, de una muestra de 20 acciones fantasma, información sobre el lugar donde se comercializa la acción —bolsa



Acción	Bolsa de cambio	Ganancia por acción ($)	Relación Precio/ganancia
Chemi-Trol	OTC	0.39	27.30
Candie's	OTC	0.07	36.20
TST/Impreso	OTC	0.65	12.70

(*continúa*)

Acción	Bolsa de cambio	Ganancia por acción	Relación precio/ganancia
Unimed Pharm.	OTC	0.12	59.30
Skyline Chili	AMEX	0.34	19.30
Cyanotech	OTC	0.22	29.30
Catalina Light.	NYSE	0.15	33.20
DDL Elect.	NYSE	0.10	10.20
Euphonix	OTC	0.09	49.70
Mesa Labs	OTC	0.37	14.40
RCM Tech.	OTC	0.47	18.60
Anuhco	AMEX	0.70	11.40
Hello Direct	OTC	0.23	21.10
Hilite Industries	OTC	0.61	7.80
Alpha Tech.	OTC	0.11	34.60
Wegener Group	OTC	0.16	24.50
U.S. Home & Garden	OTC	0.24	8.70
Chalone Wine	OTC	0.27	44.40
Eng. Support Sys.	OTC	0.89	16.70
Int. Remote Imaging	AMEX	0.86	4.70

de Nueva York (NYSE), American Stock Exchange (AMEX) o directamente (OTC)— la ganancia por acción y la relación precio/ganancia.

a. Con los datos de bolsa de cambio haga una distribución de frecuencia y otra de frecuencia relativa. ¿Cuál tiene más acciones fantasma?
b. Con los datos ganancia por acción y relación precio/ganancia elabore distribuciones de frecuencia y de frecuencia relativa. Para las ganancias por acción emplee las clases 0.00–0.19, 0.20–0.39, etc.; para la relación precio/ganancia use las clases 0.0–9.9, 10.0–19.9, etc. ¿Qué observaciones y comentarios puede hacer acerca de las acciones fantasma?

44. Los datos siguientes de la oficina de los censos de Estados Unidos proporcionan la población en millones de personas por estado (*The World Almanac,* 2006).



Population

Estado	Población	Estado	Población	Estado	Población
Alabama	4.5	Louisiana	4.5	Ohio	11.5
Alaska	0.7	Maine	1.3	Oklahoma	3.5
Arizona	5.7	Maryland	5.6	Oregon	3.6
Arkansas	2.8	Massachusetts	6.4	Pennsylvania	12.4
California	35.9	Michigan	10.1	Rhode Island	1.1
Colorado	4.6	Minnesota	5.1	South Carolina	4.2
Connecticut	3.5	Mississippi	2.9	South Dakota	0.8
Delaware	0.8	Missouri	5.8	Tennessee	5.9
Florida	17.4	Montana	0.9	Texas	22.5
Georgia	8.8	Nebraska	1.7	Utah	2.4
Hawai	1.3	Nevada	2.3	Vermont	0.6
Idaho	1.4	New Hampshire	1.3	Virginia	7.5
Illinois	12.7	New Jersey	8.7	Washington	6.2
Indiana	6.2	New Mexico	1.9	West Virginia	1.8
Iowa	3.0	New York	19.2	Wisconsin	5.5
Kansas	2.7	North Carolina	8.5	Wyoming	0.5
Kentucky	4.1	North Dakota	0.6		

a. Elabore una distribución de frecuencia, una de frecuencia porcentual y un histograma. Use como ancho de clase 2.5 millones.
b. Explique el sesgo de la distribución.
c. ¿Qué observaciones puede hacer acerca de la población en los 50 estados?

45. *Drug Store News* (septiembre de 2002) proporciona datos sobre ventas de medicamentos de las principales farmacias de Estados Unidos. Los datos siguientes son ventas anuales en millones.

Farmacia	Ventas	Farmacia	Ventas
Ahold USA	$ 1 700	Medicine Shoppe	$ 1 757
CVS	12 700	Rite-Aid	8 637
Eckerd	7 739	Safeway	2 150
Kmart	1 863	Walgreens	11 660
Kroger	3 400	Wal-Mart	7 250

a. Dé un diagrama de tallo y hojas.
b. Indique cuáles son las ventas anuales menores, mayores e intermedias.
c. ¿Cuáles son las dos farmacias mayores?

46. A continuación se presentan las temperaturas diarias más altas y más bajas registradas en 20 ciudades de Estados Unidos (*USA Today,* 3 de marzo 2006).



CityTemp

Ciudad	Alta	Baja	Ciudad	Alta	Baja
Albuquerque	66	39	Los Angeles	60	46
Atlanta	61	35	Miami	84	65
Baltimore	42	26	Minneapolis	30	11
Charlotte	60	29	New Orleans	68	50
Cincinnati	41	21	Oklahoma City	62	40
Dallas	62	47	Phoenix	77	50
Denver	60	31	Portland	54	38
Houston	70	54	St. Louis	45	27
Indianapolis	42	22	San Francisco	55	43
Las Vegas	65	43	Seattle	52	36

a. Con las temperaturas altas elabore un diagrama de tallo y hojas.
b. Con las temperaturas bajas elabore un diagrama de tallo y hojas.
c. Compare los dos diagramas y haga comentarios acerca de las diferencias entre las temperaturas más altas y las más bajas.
d. Proporcione una distribución de frecuencia de las temperaturas más altas y de las más bajas.

47. Vuelva al conjunto de datos sobre las temperaturas más altas y las temperaturas más bajas en 20 ciudades presentado en el ejercicio 46.
a. Elabore un diagrama de dispersión que muestre la relación entre las dos variables, temperatura más alta y temperatura más baja.
b. Aporte sus comentarios sobre la relación entre las temperaturas más elevadas y las más bajas.

48. Se realizó un estudio sobre satisfacción en el empleo en cuatro ocupaciones. La satisfacción en el empleo se midió mediante un cuestionario de 18 puntos en el que a cada punto había que calificarlo con una escala del 1 al 5; las puntuaciones más altas correspondían a mayor satisfacción en el empleo. La suma de las calificaciones dadas a los 18 puntos proporcionaba una medida de



OccupSat

Ocupación	Satisfacción	Ocupación	Satisfacción	Ocupación	Satisfacción
Abogado	42	Terapeuta físico	78	Analista de sistemas	60
Terapeuta físico	86	Analista de sistemas	44	Terapeuta físico	59
Abogado	42	Analista de sistemas	71	Ebanista	78
Analista de sistemas	55	Abogado	50	Terapeuta físico	60

(*continúa*)

Ocupación	Satisfacción	Ocupación	Satisfacción	Ocupación	Satisfacción
Abogado	38	Abogado	48	Terapeuta físico	50
Ebanista	79	Ebanista	69	Ebanista	79
Abogado	44	Terapeuta físico	80	Analista de sistemas	62
Analista de sistemas	41	Analista de sistemas	64	Abogado	45
Terapeuta físico	55	Terapeuta físico	55	Ebanista	84
Analista de sistemas	66	Ebanista	64	Terapeuta físico	62
Abogado	53	Ebanista	59	Analista de sistemas	73
Ebanista	65	Ebanista	54	Ebanista	60
Abogado	74	Analista de sistemas	76	Abogado	64
Terapeuta físico	52				

la satisfacción en el empleo de cada uno de los individuos de la muestra. Los datos obtenidos fueron los siguientes.

a. Dé una tabulación cruzada para ocupación y satisfacción en el trabajo.
b. En la tabulación cruzada del inciso a calcule los porcentajes de renglones.
c. ¿Qué observaciones puede hacer respecto a la satisfacción en el trabajo en estas ocupaciones?

49. ¿Generan más ingresos las grandes empresas? Los datos siguientes muestran la cantidad de empleados y el ingreso anual de 20 de las empresas de *Fortune* 1000 (*Fortune*, 17 de abril de 2000).



Empresa	Empleados	Ingreso (en millones de $)	Empresa	Empleados	Ingreso (en millones de $)
Sprint	77 600	19 930	American Financial	9 400	3 334
Chase Manhattan	74 801	33 710	Fluor	53 561	12 417
Computer Sciences	50 000	7 660	Phillips Petroleum	15 900	13 852
Wells Fargo	89 355	21 795	Cardinal Health	36 000	25 034
Sunbeam	12 200	2 398	Borders Group	23 500	2 999
CBS	29 000	7 510	MCI Worldcom	77 000	37 120
Time Warner	69 722	27 333	Consolidated Edison	14 269	7 491
Steelcase	16 200	2 743	IBP	45 000	14 075
Georgia-Pacific	57 000	17 796	Super Value	50 000	17 421
Toro	1 275	4 673	H&R Block	4 200	1 669

a. Haga un diagrama de dispersión para mostrar la relación entre las variables ingreso y empleados.
b. Haga un comentario sobre la relación entre estas variables.

50. En un sondeo realizado entre los edificios comerciales que son clientes de Cincinnati Gas & Electric Company se preguntaba cuál era el principal combustible que empleaban para la calefacción y en qué año se había construido el edificio. A continuación se presenta una parte del diagrama cruzado que se obtuvo con los datos.

Año de construcción	Tipo de combustible				
	Electricidad	Gas natural	Petróleo	Propano	Otros
1973 o antes	40	183	12	5	7
1974–1979	24	26	2	2	0
1980–1986	37	38	1	0	6
1987–1991	48	70	2	0	1

a. Complete esta tabulación cruzada dando los totales de los renglones y de las columnas.
b. Dé las distribuciones de frecuencia de año de construcción y de tipo de combustible empleado.
c. Haga una tabulación cruzada en la que se muestren los porcentajes de columnas.
d. Elabore una tabulación cruzada en la que se muestren los porcentajes de renglones.
e. Comente acerca de la relación entre año de construcción y tipo de combustible empleado.

51. La tabla 2.17 contiene parte de los datos que se encuentran en el archivo titulado Fortune en el disco compacto que viene con el libro. Este archivo proporciona fondos propios, valor de mercado y ganancias de las 50 empresas en una muestra de *Fortune 500*.

TABLA 2.17 DATOS EN UNA MUESTRA DE 50 EMPRESAS DE *FORTUNE* 500



Empresa	Fondos propios (en miles de $)	Valor de mercado (en miles de $)	Ganancias (en miles de $)
AGCO	982.1	372.1	60.6
AMP	2 698.0	12 017.6	2.0
Apple Computer	1 642.0	4 605.0	309.0
Baxter International	2 839.0	21 743.0	315.0
Bergen Brunswick	629.1	2 787.5	3.1
Best Buy	557.7	10 376.5	94.5
Charles Schwab	1 429.0	35 340.6	348.5
.	.	.	.
.	.	.	.
.	.	.	.
Walgreen	2 849.0	30 324.7	511.0
Westvaco	2 246.4	2 225.6	132.0
Whirlpool	2 001.0	3 729.4	325.0
Xerox	5 544.0	35 603.7	395.0

a. Con las variables fondos propios y ganancia elabore una tabulación cruzada. Para las ganancias emplee las clases 0–200, 200–400, …, 1000–1200. Para fondos propios emplee las clases 0–1200, 1200–2400, …, 4800–6000.
b. En la tabulación cruzada del inciso a calcule los porcentajes de renglón.
c. ¿Observa alguna relación entre ganancia y fondos propios?

52. Vuelva a la tabla 2.17.
a. Con las variables valor de mercado y ganancia elabore una tabulación cruzada.
b. En la tabulación cruzada del inciso a calcule los porcentajes de renglón.
c. Haga un comentario sobre la relación entre las variables.

53. Vuelva a la tabla 2.17.
a. Elabore un diagrama de dispersión que muestre la relación entre las variables ganancia y fondos propios.
b. Haga un comentario sobre la relación entre las variables.

54. Vuelva a la tabla 2.17.
a. Elabore un diagrama de dispersión que muestre la relación entre las variables valor de mercado y fondos propios.
b. Haga un comentario sobre la relación entre las variables.

Caso problema 1 Las tiendas Pelican

Las tiendas Pelican, una división de National Clothing, es una cadena de tiendas de ropa para mujer que tiene sucursales por todo Estados Unidos. Hace poco la tienda realizó una promoción en la que envió cupones de descuento a todos los clientes de otras tiendas de National Clothing. Los datos obtenidos en una muestra de 100 pagos con tarjeta de crédito en las tiendas Pelican durante un día de la promoción se presentan en el archivo titulado PelicanStores. En la tabla 2.18 se mues-

TABLA 2.18 DATOS DE 100 COMPRAS CON TARJETA DE CRÉDITO REALIZADAS EN LAS TIENDAS PELICAN



Cliente	Tipo de cliente	Artículos	Ventas netas	Modo de pago	Género	Estado civil	Edad
1	Regular	1	39.50	Discover	Masculino	Casado	32
2	Promocional	1	102.40	Proprietary Card	Femenino	Casada	36
3	Regular	1	22.50	Proprietary Card	Femenino	Casada	32
4	Promocional	5	100.40	Proprietary Card	Femenino	Casada	28
5	Regular	2	54.00	MasterCard	Femenino	Casada	34
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
96	Regular	1	39.50	MasterCard	Femenino	Casada	44
97	Promocional	9	253.00	Proprietary Card	Femenino	Casada	30
98	Promocional	10	287.59	Proprietary Card	Femenino	Casada	52
99	Promocional	2	47.60	Proprietary Card	Femenino	Casada	30
100	Promocional	1	28.44	Proprietary Card	Femenino	Casada	44

tra parte de este conjunto de datos. El modo de pago Propietary card se refiere a pagos realizados usando una tarjeta de crédito de National Clothing. A los clientes que hicieron compras usando un cupón de descuento se les denomina aquí promocionales y a quienes hicieron sus compras sin emplear cupón de descuento se les denomina regulares. Como a los clientes de las tiendas Pelican no se les enviaron cupones promocionales, los directivos consideran que las ventas hechas a quienes presentaron un cupón de descuento son ventas que de otro modo no se hubieran hecho. Es claro que Pelican espera que los clientes promocionales continúen comprando con ellos.

La mayor parte de las variables que aparecen en la tabla 2.18 se explican por sí mismas, pero dos de las variables deben ser aclaradas.

Artículos	El número total de artículos comprados
Ventas netas	Cantidad total cargada a la tarjeta de crédito

Los directivos de Pelican desean emplear estos datos muestrales para tener información acerca de sus clientes y para evaluar la promoción utilizando los cupones de descuento.

Informe para los directivos

Emplee los métodos tabulares y gráficos de la estadística descriptiva para ayudar a los directivos de Pelican a elaborar un perfil de sus clientes y a evaluar la promoción. Su informe debe contener, por lo menos, lo siguiente:

1. Distribuciones de frecuencia porcentual de las variables clave.
2. Una gráfica de barras o una gráfica de pastel que muestre el número de clientes correspondiente a cada modo de pago.
3. Una tabulación cruzada con el tipo de cliente (regular o promocional) frente a ventas netas. Haga un comentario sobre las semejanzas o diferencias que observe.
4. Un diagrama de dispersión para investigar la relación entre ventas netas y edad del cliente.

Caso problema 2 Industria cinematográfica

La industria cinematográfica es un negocio muy competido. En más de 50 estudios se producen de 300 a 400 películas por año y el éxito financiero de estas películas varía considerablemente. Las variables usuales para medir el éxito de una película son ventas brutas (en millones de $) en el fin de semana del estreno, ventas brutas totales (en millones de $), número de salas en que se presenta la película, semanas en las que la película se encuentra entre las 60 mejores en ventas

TABLA 2.19 DATOS DEL ÉXITO DE 10 PELÍCULAS



Película	Ventas brutas en el estreno (en millones de $)	Ventas brutas totales (en millones de $)	Número de salas	Semanas en las 60 mejores
Coach Carter	29.17	67.25	2574	16
Ladies in Lavender	0.15	6.65	119	22
Batman Begins	48.75	205.28	3858	18
Unleashed	10.90	24.47	1962	8
Pretty Persuasion	0.06	0.23	24	4
Fever Pitch	12.40	42.01	3275	14
Harry Potter and the Goblet of Fire	102.69	287.18	3858	13
Monster-in-Law	23.11	82.89	3424	16
White Noise	24.11	55.85	2279	7
Mr. and Mrs. Smith	50.34	186.22	3451	21

brutas. Los datos de una muestra de 100 películas producidas en 2005 se encuentran en el archivo titulado Movies. La tabla 2.19 muestra los datos de las 10 primeras películas que se encuentran en este archivo.

Informe para los directivos

Emplee los métodos tabulares y gráficos de la estadística descriptiva para saber cómo contribuyen estas variables al éxito de una película. Su informe debe contener lo siguiente.

1. Resúmenes tabular y gráfico de las cuatro variables interpretando cada resumen acerca de la industria cinematográfica.
2. Un diagrama de dispersión para investigar la relación entre ventas brutas totales y ventas brutas en el fin de semana del estreno. Analícelo.
3. Un diagrama de dispersión para investigar la relación entre ventas brutas totales y número de salas. Analícelo.
4. Un diagrama de dispersión para investigar la relación entre ventas brutas totales y número de semanas entre las 60 mejores. Analícelo.

Apéndice 2.1 Uso de Minitab para presentaciones gráficas y tabulares

Minitab ofrece amplias posibilidades para la elaboración de resúmenes tabulares y gráficos de datos. Minitab se usa para elaborar diversos resúmenes gráficos y tabulaciones cruzadas. Los métodos gráficos son: gráfica de puntos, histograma, diagrama de tallo y hojas y diagrama de dispersión.

Gráficas de puntos



Para esta demostración emplee los datos de la tabla 2.4 sobre las duraciones de las auditorías. Los datos se encuentran en la columna C1 de la hoja de cálculo de Minitab. Con los pasos siguientes se generará una gráfica de puntos.

Paso 1. Seleccionar el menú **Graph** y elegir **Dotplot**
Paso 2. Seleccionar **One Y, Simple** y hacer clic en **OK**
Paso 3. Cuando aparezca el cuadro de diálogo de Dotplot-One Y, Simple:
Ingresar C1 en el cuadro **Graph Variables**.
Hacer clic en **OK**

Histograma


Audit

Empleando los datos de la tabla 2.4 sobre las duraciones de las auditorías se explicará cómo se construye un histograma con las frecuencias sobre el eje vertical. Los datos están en la columna C1 de la hoja de cálculo de Minitab. Con los pasos siguientes se generará un histograma de las duraciones de las auditorías.

Paso 1. Seleccionar el menú **Graph**
Paso 2. Elegir **Histogram**
Paso 3. Seleccionar **Simple** y hacer clic en **OK**
Paso 4. Cuando aparezca el cuadro de diálogo Histogram-Simple:
Ingresar C1 en el cuadro **Graph Variables**
Hacer clic en **OK**
Paso 5. Cuando aparezca el histograma:
Posicionar el cursor del mouse sobre cualquiera de las barras
Dar doble clic
Paso 6. Cuando aparezca el cuadro de diálogo Edit Bars**:**
Hacer clic en la pestaña **Binning**
Seleccionar **Cutpoint** en Interval Type
Seleccionar **Midpoint/Cutpoint positions** en Interval Definition
Ingresar 10:35/5 en el cuadro **Midpoint/Cutpoint positions***
Hacer clic en **OK**

Observe que Minitab también proporciona la posibilidad de mostrar los puntos medios de los rectángulos del histograma como escala en el eje *x*. Si se desea esta opción, se modifica el paso 6 seleccionando **Midpoint** en Interval Definition e ingresando 12:32/5 en el cuadro **Midpoint/Cutpoint positions**. Con estos pasos se obtiene el mismo histograma pero con los puntos medios, 12, 17, 22, 27 y 32, marcados en los rectángulos del histograma.

Diagrama de tallo y hojas


ApTest

Se emplearán los datos de la tabla 2.8 sobre el examen de aptitudes para mostrar la construcción de un diagrama de tallo y hojas. Los datos se encuentran en la columna C1 de la hoja de cálculo de Minitab. Mediante los pasos siguientes se genera el diagrama extendido de tallo y hojas que se muestra en la sección 2.3.

Paso 1. Seleccionar el menú **Graph**
Paso 2. Elegir **Steam-and-Leaf**
Paso 3. Cuando aparezca el cuadro de diálogo Steam-and-Leaf:
Ingresar C1 en el cuadro **Graph Variables**
Hacer clic en **OK**

Diagrama de dispersión


Stereo

Para demostrar la elaboración de un diagrama de dispersión se emplearán los datos de la tienda de equipos de sonido que se presentan en la tabla 2.12. Las semanas están numeradas del 1 al 10 en la columna C1, los datos del número de comerciales se encuentran en la columna C2 y los datos de las ventas están en la columna C3 de la hoja de cálculo de Minitab. Con los pasos siguientes se generará el diagrama de dispersión que se muestra en la figura 2.7.

*10:35/5 indica que 10 es el valor inicial del histograma, 35 es el valor final del histograma y 5 es el ancho de clase.

Paso 1. Seleccionar el menú **Graph**
Paso 2. Elegir **Scatterplot**
Paso 3. Elegir **Simple** y dar clic en **OK**
Paso 4. Cuando aparezca el cuadro de diálogo Scatterplot-Simple:
Ingresar C3 bajo **Y variables** y C2 bajo **X variables**.
Hacer clic en **OK**

Tabulación cruzada



Restaurant

Para demostrar la elaboración de una tabulación cruzada se usan los datos de *Zagat's Restaurant Review*, parte de los cuales se muestran en la tabla 2.9. Los restaurantes se encuentran numerados del 1 al 300 en la columna C1 de la hoja de cálculo de Minitab. Los datos sobre la calidad en la columna C2 y los precios en la columna C3.

Minitab sólo puede elaborar una tabulación cruzada con variables cualitativas y el precio es una variable cuantitativa. De manera que primero necesita codificar los precios especificando la clase a la que pertenece cada precio. Con los pasos siguientes se codificarán los precios haciendo cuatro clases de precios en la columna C4: $10–19, $20–29, $30–39 y $40–49.

Paso 1. Seleccionar el menú **Data**
Paso 2. Elegir **Code**
Paso 3. Elegir **Numeric to Text**
Paso 4. Cuando aparezca el cuadro de diálogo Code-Numeric to Text:
Ingresar C3 en el cuadro **Code data from columns**
Ingresar C4 en el cuadro **Into Columns**
Ingresar 10:19 en el primer cuadro **Original values** y $10–19 en el cuadro adyacente **New**
Ingresar 20:29 en el primer cuadro **Original values** y $20–29 en el cuadro adyacente **New**
Ingresar 30:39 en el primer cuadro **Original values** y $30–39 en el cuadro adyacente **New**
Ingresar 40:49 en el primer cuadro **Original values** y $40–49 en el cuadro adyacente **New**
Hacer clic en **OK**

Para cada precio de la columna C3 aparecerá ahora su categoría correspondiente en la columna C4. Ahora puede elaborar una tabulación cruzada para calidad y categoría de los precios usando los datos de las columnas C2 y C4. Con los pasos siguientes se creará una tabulación cruzada que contendrá la misma información que la tabla 2.10.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Tables**
Paso 3. Elegir **Cross Tabulation and Chi-Square**
Paso 4. Cuando aparezcan los cuadros: Cross Tabulation y Chi-Square:
Ingresar C2 en el cuadro **For rows** y C4 en el cuadro **For columns**
Seleccionar **Counts**
Hacer clic en **OK**

Apéndice 2.2 Uso de Excel para presentaciones gráficas y tabulares

Excel ofrece amplias posibilidades para la elaboración de resúmenes tabulares y gráficos de datos. En este capítulo se muestra cómo usar Excel para elaborar una distribución de frecuencia, gráficas de barras, gráficas de pastel, histogramas, tabulaciones cruzadas y diagramas de dispersión. Se presentan dos de las herramientas más potentes de Excel: el asistente para gráficos y el informe de tabla dinámica

Distribución de frecuencia y gráficas de barras con datos cualitativos

En esta sección se muestra el uso de Excel para la elaboración de una distribución de frecuencia y de una gráfica de barras con datos cualitativos. Ambas cosas se ilustran empleando los datos de la tabla 2.1 sobre ventas de refrescos.

Distribución de frecuencia Se empezará por mostrar el uso de la función COUNTIF para elaborar una distribución de frecuencia con los datos de la tabla 2.1. Consulte la figura 2.10 a medida que se presentan los pasos de esta explicación. La hoja de cálculo con las fórmulas (en la que se ven las funciones y fórmulas empleadas) aparece en segundo plano y la hoja de cálculo con los valores (en la que aparecen los resultados obtenidos con las funciones y fórmulas usadas) aparece en primer plano.



SoftDrink

En las celdas A1:A51 se encuentra el título "Ventas de refrescos" y los datos de 50 ventas de refrescos. En las celdas C1:D1 también se ingresaron los títulos "Refresco" y "Frecuencia". Los nombres de los cinco refrescos se ingresaron en las celdas C2:C6. Ahora se puede usar la función COUNTIF de Excel para contar cuántas veces aparece cada refresco en las celdas A2:A51. Para esto se siguen los pasos:

Paso 1. Seleccionar la celda D2
Paso 2. Ingresar =COUNTIF($A$2:$A$51,C2)
Paso 3. Copiar la celda D2 a las celdas D3:D6

FIGURA 2.10 DISTRIBUCIÓN DE FRECUENCIA DE LAS VENTAS DE REFRESCOS CONSTRUIDA EMPLEANDO LA FUNCIÓN COUNTIF DE EXCEL

	A	B	C	D	E
1	**Ventas de refrescos**		**Refresco**	**Frecuencia**	
2	Coke Classic		Coke Classic	=COUNTIF($A$2:$A$51,C2)	
3	Diet Coke		Diet Coke	=COUNTIF($A$2:$A$51,C3)	
4	Pepsi		Dr. Pepper	=COUNTIF($A$2:$A$51,C4)	
5	Diet Coke		Pepsi	=COUNTIF($A$2:$A$51,C5)	
6	Coke Classic		Sprite	=COUNTIF($A$2:$A$51,C6)	
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

	A	B	C	D	E
1	**Ventas de refrescos**		**Refresco**	**Frecuencia**	
2	Coke Classic		Coke Classic	19	
3	Diet Coke		Diet Coke	8	
4	Pepsi		Dr. Pepper	5	
5	Diet Coke		Pepsi	13	
6	Coke Classic		Sprite	5	
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

Nota: Los renglones 11–44 están ocultos.

En la hoja de cálculo con las fórmulas de la figura 2.10 se observan en las celdas las fórmulas ingresadas al seguir estos pasos. En la hoja de cálculo con los valores se observan los valores obtenidos con las fórmulas de cada celda. En esta hoja de cálculo se aprecia la misma distribución de frecuencia de la tabla 2.2



SoftDrink

Gráfica de barras Aquí se muestra cómo usar el asistente para gráficos de Excel para elaborar una gráfica de barras con los datos de las ventas de refrescos. En la figura 2.10 obsérvese la distribución de frecuencia que se presenta en la hoja de cálculo con los valores. La gráfica de barras que se va a construir es una extensión de esta hoja de cálculo. En la figura 2.11 se muestra esta misma hoja de cálculo con la gráfica de barras elaborada usando el asistente para gráficas. Los pasos a seguir son:

Paso 1. Seleccionar las celdas C1:D6

Paso 2. Hacer clic en el botón **Asistente para gráficas** de la barra de herramientas estándar (o seleccionar el menú **Insertar** y elegir la opción **Gráfico**)

Paso 3. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 1 de 4: tipo de gráfico:

Elegir **Columnas** de la lista **Tipo de gráfico**
Elegir **Columnas agrupadas** en la visualización **Subtipo de gráfico**
Hacer clic en **Siguiente >**

Paso 4. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 2 de 4: datos de origen:

Hacer clic en **Siguiente >**

FIGURA 2.11 GRÁFICA DE BARRAS CON LOS DATOS DE LAS VENTAS DE REFRESCOS ELABORADA MEDIANTE EL ASISTENTE PARA GRÁFICOS DE EXCEL

	A	B	C	D	E	F	G	H	I
1	**Ventas de refrescos**		**Refresco**	**Frecuencia**					
2	Coke Classic		Coke Classic	19					
3	Diet Coke		Diet Coke	8					
4	Pepsi		Dr. Pepper	5					
5	Diet Coke		Pepsi	13					
6	Coke Classic		Sprite	5					
7	Coke Classic								
8	Dr. Pepper								
9	Diet Coke								
10	Pepsi								
45	Pepsi								
46	Pepsi								
47	Pepsi								
48	Coke Classic								
49	Dr. Pepper								
50	Pepsi								
51	Sprite								
52									
53									
54									
55									
56									
57									

Paso 5. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 3 de 4: opciones de gráfico:
Seleccionar la pestaña **Títulos** y después
Digitar Gráfica de barras con los datos de las ventas de refrescos en el cuadro **Título del gráfico**
Digitar Refresco en el cuadro **Eje de categorías (X)**
Digitar Frecuencia en el cuadro **Eje de valores (Y)**
Seleccionar la pestaña **Leyenda** y después
Quitar la paloma (marca de verificación) que aparece en el cuadro **Mostrar leyenda**
Hacer clic en **Siguiente >**

Paso 6. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 4 de 4: ubicación:
Especificar una ubicación para la nueva gráfica (aquí se utilizó la misma hoja de cálculo que se estaba empleando por lo que se eligió la opción **Como objeto en**)
Hacer clic en **Finalizar**

En la figura 2.11 se muestra la gráfica de barras que se obtuvo.*

De manera similar, Excel puede elaborar una gráfica de pastel con los datos de las ventas de refrescos. La diferencia principal es que en el paso 3 se elige **Circular** de la lista Tipo de gráfico.

Distribuciones de frecuencia e histogramas para datos cuantitativos

En esta sección se muestra cómo usar Excel para elaborar una distribución de frecuencia y un histograma con datos cuantitativos. Para ilustrar esto se usan los datos de la tabla 2.4 sobre la duración de las auditorías.

archivo en CD
Audit

Distribución de frecuencia Para elaborar una distribución de frecuencia con datos cuantitativos se puede usar la función FREQUENCY de Excel. Consulte la figura 2.12 a medida que se presentan los pasos a seguir. La hoja de cálculo con las fórmulas aparece en segundo plano y la hoja de cálculo con los valores aparece en primer plano. El título “Duración de la auditoría” se encuentra en la celda A1 y los datos de las 20 auditorías están en las celdas A2:A21. Siguiendo los procedimientos indicados en el texto, introduzca las cinco clases 10–14, 15–19, 20–24, 25–29 y 30–34. El título “Duración de la auditoría” y las cinco clases se ingresan en las celdas C1:C6. El título “Límite superior” y los cinco límites superiores de las clases se ingresan en las celdas D1:D6. Ingrese también el título “Frecuencia” en la celda E1. La función FREQUENCY de Excel se usará para obtener la frecuencia en las celdas E2:E6. Los pasos siguientes describen cómo elaborar una distribución de frecuencia con los datos de las duraciones de las auditorías.

Para ingresar una fórmula matricial es necesario mantener oprimidas las teclas Ctrl y Shift(mayúsculas) mientras se pulsa la tecla Enter.

Paso 1. Seleccionar las celdas E2:E6

Paso 2. Digitar, pero no ingresar, la fórmula siguiente:

=FREQUENCY(A2:A21,D2:D6)

Paso 3. Pulsar las teclas CTRL+SHIFT(mayúsculas)+ENTER con lo que la fórmula matricial será ingresada en cada una de las celdas E2:E6

El resultado se muestra en la figura 2.12. Los valores que aparecen en las celdas E2:E6 son las frecuencias de las clases correspondientes. Regrese a la función FREQUENCY, vea que el intervalo de las celdas para los límites superiores de clase (D2:D6) sirve de argumento a la función. Estos límites superiores de clase a los que Excel llama *bins*, le dicen a Excel qué frecuencia poner en las celdas del intervalo de salida (E2:E6). Por ejemplo, la frecuencia de la clase que tiene el límite superior, o *bin*, 14 será colocada en la primera celda (E2), la frecuencia de la clase que tiene el límite superior, o *bin*, 19 será colocada en la segunda celda (E3), y así sucesivamente.

*La gráfica de barras de la figura 2.11 no es del mismo tamaño que la obtenida con Excel después de seleccionar **Finalizar**. Modificar el tamaño de una gráfica de Excel no es difícil. Primero se selecciona la gráfica, en los bordes de la gráfica aparecerán unos cuadritos negros llamados manillas de tamaño. Hacer clic sobre las manillas de tamaño y arrastrarlas para darle a la figura el tamaño deseado.

FIGURA 2.12 DISTRIBUCIÓN DE FRECUENCIA DE LOS DATOS DE LAS DURACIONES DE LAS AUDITORÍAS CON LA FUNCIÓN FREQUENCY DE EXCEL

	A	B	C	D	E
1	D. auditoría		D. auditoría	Límite superior	Frecuencia
2	12		10-14	14	=FREQUENCY(A2:A21,D2:D6)
3	15		15-19	19	=FREQUENCY(A2:A21,D2:D6)
4	20		20-24	24	=FREQUENCY(A2:A21,D2:D6)
5	22		25-29	29	=FREQUENCY(A2:A21,D2:D6)
6	14		30-34	34	=FREQUENCY(A2:A21,D2:D6)
7	14				
8	15				
9	27				
10	21				
11	18				
12	19				
13	18				
14	22				
15	33				
16	16				
17	18				
18	17				
19	23				
20	28				
21	13				

	A	B	C	D	E
1	D. auditoría		D. auditoría	Límite superior	Frecuencia
2	12		10-14	14	4
3	15		15-19	19	8
4	20		20-24	24	5
5	22		25-29	29	2
6	14		30-34	34	1
7	14				
8	15				
9	27				
10	21				
11	18				
12	19				
13	18				
14	22				
15	33				
16	16				
17	18				
18	17				
19	23				
20	28				
21	13				

Histograma Para usar el ayudante para gráficos de Excel para construir un histograma con las duraciones de las auditorías parta de la distribución de frecuencia de la figura 2.12. En la figura 2.13 se presenta la hoja de trabajo con la distribución de frecuencia y el histograma. Los pasos siguientes indican cómo emplear el asistente para gráficos al elaborar un histograma con los datos de las duraciones de las auditorías.

Paso 1. Seleccionar las celdas E2:E6

Paso 2. Hacer clic en el botón **Asistente para gráficas** de la barra de herramientas estándar (o seleccionar el menú **Insertar** y elegir la opción **Gráfico**)

Paso 3. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 1 de 4: tipo de gráfico

Elegir **Columnas** en la lista **Tipo de gráfico**
Elegir **Columnas agrupadas** en la visualización **Subtipo de gráfico**
Hacer clic en **Siguiente >**

Paso 4. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 2 de 4: datos de origen:

Seleccionar la pestaña **Serie** y después
Hacer clic en el cuadro **Rótulos del eje de categorías (X)**
Seleccionar las celdas C2:C6
Hacer clic en **Siguiente >**

FIGURA 2.13 HISTOGRAMA CON LAS DURACIONES DE LAS AUDITORÍAS

	A	B	C	D	E	F	G
1	D. auditoría		D. auditoría	Límite superior	Frecuencia		
2	12		10-14	14	4		
3	15		15-19	19	8		
4	20		20-24	24	5		
5	22		25-29	29	2		
6	14		30-34	34	1		
7	14						
8	15						
9	27						
10	21						
11	18						
12	19						
13	18						
14	22						
15	33						
16	16						
17	18						
18	17						
19	23						
20	28						
21	13						
22							

Paso 5. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 3 de 4: opciones de gráfico:
Seleccionar la pestaña **Títulos** y después
Digitar Histograma de las duraciones de las auditorías en el cuadro **Título del gráfico**
Digitar Duración de las auditorías en días en el cuadro **Eje de categorías (X)**:
Digitar Frecuencia en el cuadro **Eje de valores (Y)**:
Seleccionar la pestaña **Leyenda** y después
Quitar la paloma (marca de verificación) que aparece en el cuadro **Mostrar leyenda**
Hacer clic en **Siguiente >**

Paso 6. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 4 de 4: ubicación:
Especificar una ubicación para la nueva gráfica (aquí se utilizó la misma hoja de cálculo que se estaba empleando por lo que se eligió la opción **Como objeto en**)
Hacer clic en **Finalizar**

Ahora en la hoja de cálculo aparecerá una gráfica de columnas elaborada por Excel. Pero entre las columnas habrá espacios. Como en un histograma no hay espacios entre las columnas, es necesario modificar esta gráfica para eliminar los espacios entre las columnas. Los pasos siguientes describen cómo hacerlo.

Paso 1. Dar doble clic en cualquiera de las columnas de la gráfica.

Paso 2. Cuando aparezca el cuadro de diálogo Formato de punto de datos:
Seleccionar la pestaña **Opciones**
Ingresar 0 en el cuadro **Ancho del rango**
Hacer clic en **Aceptar**

El histograma se verá como el que aparece en la figura 2.13.

Por último, un aspecto interesante de la hoja de cálculo de la figura 2.13 es que Excel ha relacionado los datos que aparecen en las celdas A2:A21 con las frecuencias que aparecen en las celdas E2:E6 y con el histograma. Si se modifica alguno de los datos de las celdas A2:A21 se

FIGURA 2.14 DIAGRAMA DE DISPERSIÓN DE LAS VENTAS DE LA TIENDA DE EQUIPOS DE SONIDO

	A	B	C	D	E	F	G	H
1	Semana	Cantidad de comerciales	Volumen de ventas					
2	1	2	50					
3	2	5	57					
4	3	1	41					
5	4	3	54					
6	5	4	54					
7	6	1						
8	7	5						
9	8	3						
10	9	4						
11	10	2						
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

modificarán automáticamente las frecuencias de las celdas E2:E6 y también el histograma y aparecerán una distribución de frecuencias y un histograma modificados. Se aconseja probar cómo se realiza esta modificación automática modificando uno o dos de los datos.

Diagrama de dispersión



Stereo

Se usarán los datos de la tienda de equipo de sonido que aparecen en la tabla 2.12 para mostrar cómo se usa el asistente para gráficos de Excel al elaborar un diagrama de dispersión. Consulte la figura 2.14 a medida que se describen los pasos para elaborar esta gráfica. La hoja de cálculo con los valores aparece en segundo plano y el diagrama de dispersión elaborado por el asistente para gráficos en primer plano. Los pasos a seguir son los siguientes.

Paso 1. Seleccionar las celda B1:C11

Paso 2. Hacer clic en el botón **Asistente para gráficas** de la barra de herramientas estándar (o seleccionar el menú **Insertar** y elegir la opción **Gráfico**)

Paso 3. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 1 de 4: tipo de gráfico:

Elegir **XY** (**Dispersión**) en la lista **Tipo de gráfico**
Elegir **Dispersión** en la visualización **Subtipo de gráfico**
Hacer clic en **Siguiente >**

Paso 4. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 2 de 4: datos de origen:

Hacer clic en **Siguiente >**

Paso 5. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 3 de 4: opciones de gráfico:

Seleccionar la pestaña **Títulos**
Digitar Diagrama de dispersión de las ventas de la tienda de equipos de sonido en el cuadro **Título del gráfico**
Digitar Número de comerciales en el cuadro **Eje de categorías (X)**:
Digitar Volumen de ventas en el cuadro **Eje de valores (Y)**:

Seleccionar la pestaña **Leyenda**
Quitar la paloma (marca de verificación) que aparece en el cuadro **Mostrar leyenda**
Hacer clic en **Siguiente >**

Paso 6. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 4 de 4: ubicación:
Especificar una ubicación para la nueva gráfica (aquí se utilizó la misma hoja de cálculo que se estaba empleando por lo que se eligió la opción **Como objeto en**)
Hacer clic en **Finalizar**

En el diagrama de dispersión puede trazar una línea de tendencia de la manera siguiente.

Paso 1. Colocar el cursor del mouse sobre cualquiera de los puntos del diagrama de dispersión y dar clic con el botón derecho del mouse. Aparecerá una lista de opciones
Paso 2. Elegir **Agregar línea de tendencia**
Paso 3. Cuando aparezca el cuadro agregar línea de tendencia:
Seleccionar la pestaña **Tipo**
Elegir **Lineal** en la visualización **Tipo de tendencia o regresión**
Hacer clic en **Aceptar**

En la hoja de cálculo de la figura 2.14 se observa el diagrama de dispersión con la línea de tendencia.

Informe en tabla dinámica

El informe en tabla dinámica de Excel proporciona una valiosa herramienta para la manipulación de un conjunto de datos en que se tiene más de una variable. Se ilustrará su uso mostrando cómo elaborar una tabulación cruzada.

Tabulación cruzada Se ilustra la elaboración de una tabulación cruzada empleando los datos de los restaurantes que aparecen en la figura 2.15. Los títulos se han ingresado en el renglón 1 y los datos de los 300 restaurantes se han ingresado en las celdas A2:C301

FIGURA 2.15 HOJA DE CÁLCULO DE EXCEL CON LOS DATOS DE LOS RESTAURANTES



Nota: los renglones 12–291 están ocultos.

	A	B	C	D
1	**Restaurante**	**Calidad**	**Precio ($)**	
2	1	Bueno	18	
3	2	Muy bueno	22	
4	3	Bueno	28	
5	4	Excelente	38	
6	5	Muy bueno	33	
7	6	Bueno	28	
8	7	Muy bueno	19	
9	8	Muy bueno	11	
10	9	Muy bueno	23	
11	10	Bueno	13	
292	291	Muy bueno	23	
293	292	Muy bueno	24	
294	293	Excelente	45	
295	294	Bueno	14	
296	295	Bueno	18	
297	296	Bueno	17	
298	297	Bueno	16	
299	298	Bueno	15	
300	299	Muy bueno	38	
301	300	Muy bueno	31	
302				

Paso 1. Seleccionar el menú **Datos**
Paso 2. Elegir **Informe de tabla y datos dinámicos**
Paso 3. Cuando aparezca el cuadro de diálogo Asistente para tablas y gráficos dinámicos – paso 1 de 3:
Elegir **Lista o base de datos de Microsoft Excel**
Elegir **Tabla dinámica**
Hacer clic en **Siguiente**
Paso 4. Cuando aparezca el cuadro de diálogo Asistente para tablas y gráficos dinámicos – paso 2 de 3:
Ingresar A1:C301 en el cuadro **Rango**
Hacer clic en **Siguiente**
Paso 5. Cuando aparezca el cuadro de diálogo Asistente para tablas y gráficos dinámicos – paso 3 de 3:
Seleccionar **Hoja de cálculo nueva**
Seleccionar **Diseño**
Paso 6. Cuando aparezca el diagrama Asistente para tablas y gráficos dinámicos – diseño (véase figura 2.16):
Arrastre el botón de campo **Calidad (Quality)** a la sección **FILA (ROW)** del diagrama
Arrastre el botón de campo **Precio (Meal Price)** a la sección **COLUMNA (COLUMN)** del diagrama
Arrastre el botón de campo **Restaurante (Restaurant)** a la sección **DATOS (DATA)** del diagrama
Dar doble clic en el botón de campo **Suma de Restaurante** en la sección DATOS
Cuando aparezca el cuadro de diálogo Campo de la tabla dinámica:
Elegir **Cuenta** bajo **Resumir por**
Hacer clic en **Aceptar** (la figura 2.17 muestra el diseño completo del diagrama)
Hacer clic en **Aceptar**
Paso 7. Cuando aparezca el cuadro de diálogo Asistente para tablas y gráficos dinámicos – paso 3 de 3:
Hacer clic en **Finalizar**

En la figura 2.18 se muestra parte del resultado generado por Excel. Observe que las columnas D a AK se han ocultado para que se puedan mostrar los resultados en una figura de tamaño razo-

FIGURA 2.16 ASISTENTE PARA TABLAS Y GRÁFICOS DINÁMICOS: DISEÑO

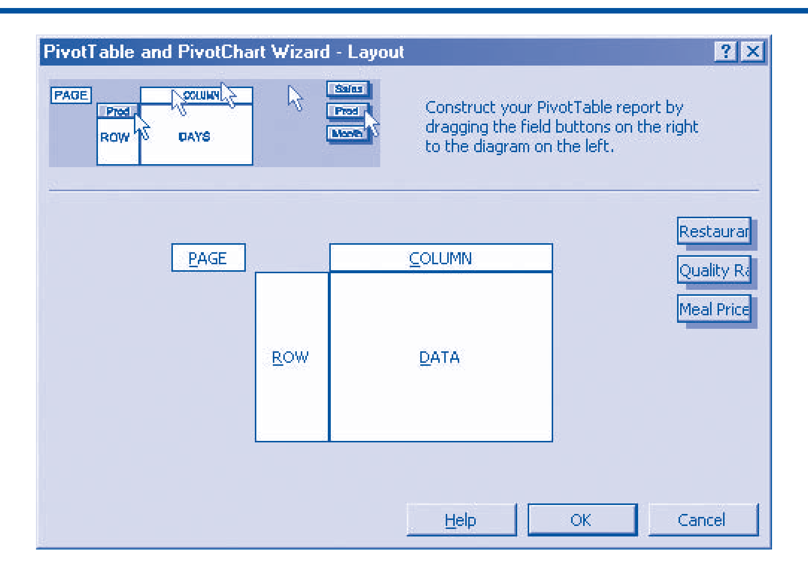


FIGURA 2.17 ASISTENTE PARA TABLAS Y GRÁFICOS DINÁMICOS: DISEÑO

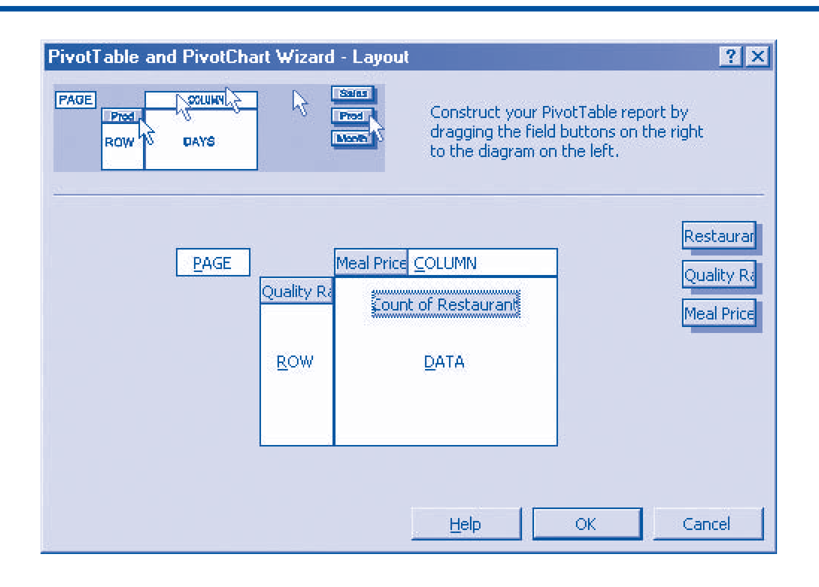


FIGURA 2.18 RESULTADO INICIAL DEL INFORME DE TABLA DINÁMICA (LAS COLUMNAS D:AK ESTÁN OCULTAS)

	A	B	C	AL	AM	AN	AO
1							
2							
3	Suma de restaurantes	Precio ($) ▼					
4	Calidad ▼	10	11	47	48	Gran total	
5	Excelente			2	2	66	
6	Bueno	6	4			84	
7	Muy bueno	1	4		1	150	
8	Gran total	7	8	2	3	300	
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							

FIGURA 2.19 INFORME DE TABLA DINÁMICA FINAL CON LOS DATOS DE LOS RESTAURANTES

	A	B	C	D	E	F	G
1							
2							
3	Suma de restaurantes	Precio ($) ▼					
4	Calidad ▼	10-19	20-29	30-39	40-49	Gran total	
5	Bueno	42	40	2		84	
6	Muy bueno	34	64	46	6	150	
7	Excelente	2	14	28	22	66	
8	Gran total	78	118	76	28	300	
9							
10							
11							
12		PivotTable					
13		PivotTable ▾					
14							
15							
16							
17							
18							
19							
20							

nable. Los títulos de los renglones (Excelente, Bueno y Muy bueno) y los totales de los renglones (66, 84, 150 y 300) de la figura 2.18 son los mismos que en la tabla 2.10, sólo que en distinto orden. Para colocarlos en el orden Bueno, Muy bueno, Excelente hay que seguir los siguientes pasos:

Paso 1. Hacer clic con el botón derecho sobre la celda A5
Paso 2. Elegir **Ordenar**
Paso 3. Elegir **Mover al final**

En la figura 2.18 hay una columna para cada precio. Por ejemplo, en la columna B se encuentran los restaurantes cuyo precio es $10, en la columna C los restaurantes cuyo precio es $11, etc. Para que el informe en tabla dinámica se vea como en la tabla 2.10, se deben agrupar las columnas en cuatro categorías de precios: $10–19, $20–29, $30–39 y $40–49. Los pasos necesarios para agrupar las columnas de la hoja de cálculo que aparece en la figura 2.18 son:

Paso 1. Hacer clic con el botón derecho en Precio($) en la celda B3 de la Tabla dinámica
Paso 2. Elegir **Agrupar y mostrar detalles**
Elegir **Agrupar**
Paso 3. Cuando aparezca el cuadro de diálogo **Agrupar**
Ingresar 10 en el cuadro **Comenzar en**
Ingresar 49 en el cuadro **Terminar en**
Ingresar 10 en el cuadro **Por**
Hacer clic en **Aceptar**

La tabla dinámica que se obtiene se presenta en la figura 2.19. Es la tabla dinámica final. Observe que esta tabla proporciona la misma información que la tabla cruzada de la tabla 2.10.

CAPÍTULO 3



Estadística descriptiva: medidas numéricas

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

SMALL FRY DESIGN*
SANTA ANA, CALIFORNIA

Fundada en 1997, Small Fry Design es una empresa de juguetes y accesorios que diseña e importa productos para niños pequeños. La línea de productos de la empresa incluye muñecos de peluche, móviles, juguetes musicales, sonajeros y mantas de seguridad y ofrece diseños de juguetes de alta calidad para bebés, con énfasis especial en los colores, texturas y sonidos. Los productos son diseñados en Estados Unidos y manufacturados en China.

Small Fry Design emplea representantes independientes para la venta de sus productos a tiendas de mobiliario para niños, tiendas de accesorios y ropa para niños, tiendas de regalos, tiendas exclusivas de departamentos e importantes empresas de ventas por catálogo. En la actualidad los productos de Small Fry Design se distribuyen en más de 1000 negocios en todo Estados Unidos.

La administración del flujo de efectivo es una de las actividades más relevantes del funcionamiento cotidiano de esta empresa. Garantizar suficiente ingreso de efectivo para cumplir con la deuda corriente y la deuda a corto plazo es la diferencia entre el éxito y el fracaso de la empresa. Un factor importante de la administración del flujo de efectivo es el análisis y control de las cuentas por cobrar. Al medir el tiempo promedio y el valor en dólares que tienen las facturas pendientes, los administradores pronostican la disponibilidad de dinero y vigilan la situación de las cuentas por cobrar. La empresa se ha planteado los objetivos siguientes: el tiempo promedio de una factura pendiente no debe ser más de 45 días y el valor en dólares de las facturas que tengan más de 60 días no debe ser superior a 5% del valor en dólares de todas las cuentas por cobrar.

En un resumen reciente sobre el estado de las cuentas por cobrar se presentaron los siguientes estadísticos descriptivos sobre el tiempo que tenían las facturas pendientes.

Media	40 días
Mediana	35 días
Moda	31 días

*Los autores agradecen a John A. McCarthy, presidente de Small Fry Design por proporcionar este artículo para *La estadística en la práctica.*



Móvil “El rey de la selva” de Small Fry Design.
© Foto cortesía de Small Fry Design, Inc.

La interpretación de dichos estadísticos indica que el tiempo promedio de una factura pendiente es 40 días. La mediana revela que la mitad de las facturas se quedan pendientes 35 días o más. La moda, 31 días, muestra que el tiempo que con más frecuencia permanece pendiente una factura es 31 días. Este resumen estadístico indica también que sólo 3% del valor en dólares de todas las cuentas por cobrar tienen más de 60 días. De acuerdo con esta información estadística, la administración está satisfecha de que las cuentas por cobrar y el flujo de efectivo entrante estén bajo control.

En este capítulo aprenderá a calcular e interpretar algunas de las medidas estadísticas empleadas por Small Fry Design. Además de la media, la mediana y la moda usted estudiará otros estadísticos descriptivos como el rango, la varianza, la desviación estándar, los percentiles y la correlación. Estas medidas numéricas ayudan a la comprensión e interpretación de datos.

En el capítulo 2 estudió las presentaciones tabular y gráfica para resumir datos. En este capítulo se le presentan varias medidas numéricas que proporcionan otras opciones para resumir datos.

Empezará con medidas numéricas para conjuntos de datos que constan de una sola variable. Si el conjunto de datos consta de más de una variable, empleará estas mismas medidas numéricas para cada una de las variables por separado. Sin embargo, en el caso de dos variables, estudiará también medidas de la relación entre dos variables.

Se presentan medidas numéricas de localización, dispersión, forma, y asociación. Si estas medidas las calcula con los datos de una muestra, se llaman **estadísticos muestrales**. Si estas medidas las calcula con los datos de una población se llaman **parámetros poblacionales**. En inferencia estadística, al estadístico muestral se le conoce como el **estimador puntual** del correspondiente parámetro poblacional. El proceso de estimación puntual será estudiado con más detalle en el capítulo 7.

En los dos apéndices del capítulo se le muestra cómo usar Minitab y Excel para calcular muchas de las medidas descritas en este capítulo.

3.1 Medidas de localización

Media

La medida de localización más importante es la **media**, o valor promedio, de una variable. La media proporciona una medida de localización central de los datos. Si los datos son datos de una muestra, la media se denota $\bar{x}$; si los datos son datos de una población, la media se denota con la letra griega μ.

En las fórmulas estadísticas se acostumbra denotar el valor de la primera observación de la variable x con x_1, el valor de la segunda observación de la variable x con x_2 y así con lo siguiente. En general, el valor de la i-ésima observación de la variable x se denota x_i. La fórmula para la media muestral cuando se tiene una muestra de n observaciones es la siguiente.

La media muestral $\bar{x}$ es un estadístico muestral.

MEDIA MUESTRAL

$$\bar{x} = \frac{\Sigma x_i}{n} \tag{3.1}$$

En la fórmula anterior el numerador es la suma de los valores de las n observaciones. Es decir,

$$\Sigma x_i = x_1 + x_2 + \cdots + x_n$$

La letra griega Σ es el símbolo de sumatoria (suma)

Para ilustrar el cálculo de la media muestral, considere los siguientes datos que representan el tamaño de cinco grupos de una universidad.

46 54 42 46 32

Se emplea la notación x_1, x_2, x_3, x_4, x_5 para representar el número de estudiantes en cada uno de los cinco grupos.

$$x_1 = 46 \qquad x_2 = 54 \qquad x_3 = 42 \qquad x_4 = 46 \qquad x_5 = 32$$

Por tanto, para calcular la media muestral, escriba

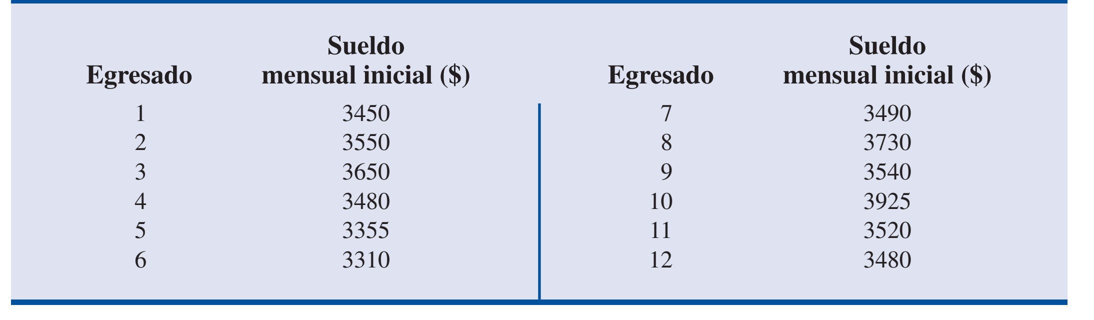
$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 54 + 42 + 46 + 32}{5} = 44$$

La media muestral del tamaño de estos grupos es 44 alumnos.

Otra ilustración del cálculo de la media muestral aparece en la situación siguiente. Suponga que la bolsa de trabajo de una universidad envía cuestionarios a los recién egresados de la carrera de administración solicitándoles información sobre sus sueldos mensuales iniciales. En la ta-



TABLA 3.1 SUELDOS MENSUALES INICIALES EN UNA MUESTRA DE 12 RECIÉN EGRESADOS DE LA CARRERA DE ADMINISTRACIÓN



Egresado	Sueldo mensual inicial ($)	Egresado	Sueldo mensual inicial ($)
1	3450	7	3490
2	3550	8	3730
3	3650	9	3540
4	3480	10	3925
5	3355	11	3520
6	3310	12	3480

bla 3.1 se presentan estos datos. El sueldo mensual inicial medio de los 12 recién egresados se calcula como sigue.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{12}}{12}$$

$$= \frac{3450 + 3550 + \cdots + 3480}{12}$$

$$= \frac{42\,480}{12} = 3540$$

En la ecuación (3.1) se muestra cómo se calcula la media en una muestra de n observaciones. Para calcular la media de una población use la misma fórmula, pero con una notación diferente para indicar que trabaja con toda la población. El número de observaciones en una población se denota N y el símbolo para la media poblacional es μ.

La media muestral $\bar{x}$ es un estimador puntual de la media poblacional μ.

MEDIA POBLACIONAL

$$\mu = \frac{\Sigma x_i}{N} \quad \textbf{(3.2)}$$

Mediana

La **mediana** es otra medida de localización central. Es el valor de enmedio en los datos ordenados de menor a mayor (en forma ascendente). Cuando tiene un número impar de observaciones, la mediana es el valor de enmedio. Cuando la cantidad de observaciones es par, no hay un número enmedio. En este caso, se sigue una convención y la mediana es definida como el promedio de las dos observaciones de enmedio. Por conveniencia, la definición de mediana se replantea así:

MEDIANA

Ordenar los datos de menor a mayor (en forma ascendente).

a. Si el número de observaciones es impar, la mediana es el valor de enmedio.
b. Si el número de observaciones es par, la mediana es el promedio de las dos observaciones de enmedio.

Apliquemos esta definición para calcular la mediana del número de alumnos en un grupo a partir de la muestra de los cinco grupos de universidad. Los datos en orden ascendente son

32 42 46 46 54

Como $n = 5$ es impar, la mediana es el valor de enmedio. De manera que la mediana del tamaño de los grupos es 46. Aun cuando en este conjunto de datos hay dos observaciones cuyo valor es 46, al poner las observaciones en orden ascendente se toman en consideración todas las observaciones.

Suponga que también desea calcular la mediana del salario inicial de los 12 recién egresados de la carrera de administración de la tabla 3.1. Primero ordena los datos de menor a mayor

$$3310 \quad 3355 \quad 3450 \quad 3480 \quad 3480 \quad \underbrace{3490 \quad 3520}_{\text{Los dos valores de en medio}} \quad 3540 \quad 3550 \quad 3650 \quad 3730 \quad 3925$$

Como $n = 12$ es par, se localizan los dos valores de enmedio: 3490 y 3520. La mediana es el promedio de estos dos valores.

$$\text{Mediana} = \frac{3490 + 3520}{2} = 3505$$

La mediana es la medida de localización más empleada cuando se trata de ingresos anuales y valores de propiedades, debido a que la media puede inflarse por unos cuantos ingresos o valores de propiedades muy altos. En tales casos, la mediana es la medida de localización central preferida.

Aunque la media es la medida de localización central más empleada, en algunas situaciones se prefiere la mediana. A la media la influyen datos en extremo pequeños o considerablemente grandes. Por ejemplo, suponga que uno de los recién graduados de la tabla 3.1 tuviera un salario inicial de $10 000 mensuales (quizá su familia sea la dueña de la empresa). Si reemplaza el mayor sueldo inicial mensual de la tabla 3.1, $3925, por $10 000 y vuelve a calcular la media, la media muestral cambia de $3540 a $4046. Sin embargo, la mediana, $3505, permanece igual ya que $3490 y $3520 siguen siendo los dos valores de en medio. Si hay algunos sueldos demasiado altos, la mediana proporciona una medida de tendencia central mejor que la media. Al generalizar lo anterior, es posible decir que cuando los datos contengan valores extremos, es preferible usar a la mediana como medida de localización central.

Moda

La tercera medida de localización es la **moda**. La moda se define como sigue.

> MODA
>
> La moda es el valor que se presenta con mayor frecuencia.

Para ilustrar cómo identificar a la moda, considere la muestra del tamaño de los cinco grupos de la universidad. El único valor que se presenta más de una vez es el 46. La frecuencia con que se presenta este valor es 2, por lo que es el valor con mayor frecuencia, entonces es la moda. Para ver otro ejemplo, considere la muestra de los sueldos iniciales de los recién egresados de la carrera de administración. El único salario mensual inicial que se presenta más de una vez es $3480. Como este valor tiene la frecuencia mayor, es la moda.

Hay situaciones en que la frecuencia mayor se presenta con dos o más valores distintos. Cuando esto ocurre hay más de una moda. Si los datos contienen más de una moda se dice que los datos son *bimodales.* Si contienen más de dos modas, son *multimodales.* En los casos multimodales casi nunca se da la moda, porque dar tres o más modas no resulta de mucha ayuda para describir la localización de los datos.

Percentiles

Un **percentil** aporta información acerca de la dispersión de los datos en el intervalo que va del menor al mayor valor de los datos. En los conjuntos de datos que no tienen muchos valores repetidos, el percentil p divide a los datos en dos partes. Cerca de p por ciento de las observaciones tienen valores menores que el percentil p y aproximadamente $(100 - p)$ por ciento de las observaciones tienen valores mayores que el percentil p. El percentil p se define como sigue:

PERCENTIL

El percentil p es un valor tal que por lo menos p por ciento de las observaciones son menores o iguales que este valor y por lo menos $(100 - p)$ por ciento de las observaciones son mayores o iguales que este valor.

Las puntuaciones en los exámenes de admisión de escuelas y universidades se suelen dar en términos de percentiles. Por ejemplo, suponga que un estudiante obtiene 54 puntos en la parte verbal del examen de admisión. Esto no dice mucho acerca de este estudiante en relación con los demás estudiantes que realizaron el examen. Sin embargo, si esta puntuación corresponde al percentil 70, entonces 70% de los estudiantes obtuvieron una puntuación menor a la de dicho estudiante y 30% de los estudiantes obtuvieron una puntuación mayor.

Para calcular el percentil p se emplea el procedimiento siguiente.

Seguir estos pasos facilita el cálculo de los percentiles.

CÁLCULO DEL PERCENTIL p

Paso 1. Ordenar los datos de menor a mayor (colocar los datos en orden ascendente).
Paso 2. Calcular el índice i

$$i = \left(\frac{p}{100}\right)n$$

donde p es el percentil deseado y n es el número de observaciones.
Paso 3. (a) Si i *no es un número entero, debe redondearlo.* El primer entero mayor que i denota la posición del percentil p.
(b) Si i *es un número entero,* el percentil p es el promedio de los valores en las posiciones i e $i + 1$.

Para ilustrar el empleo de este procedimiento, determine el percentil 85 en los sueldos mensuales iniciales de la tabla 3.1.

Paso 1. Ordenar los datos de menor a mayor

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

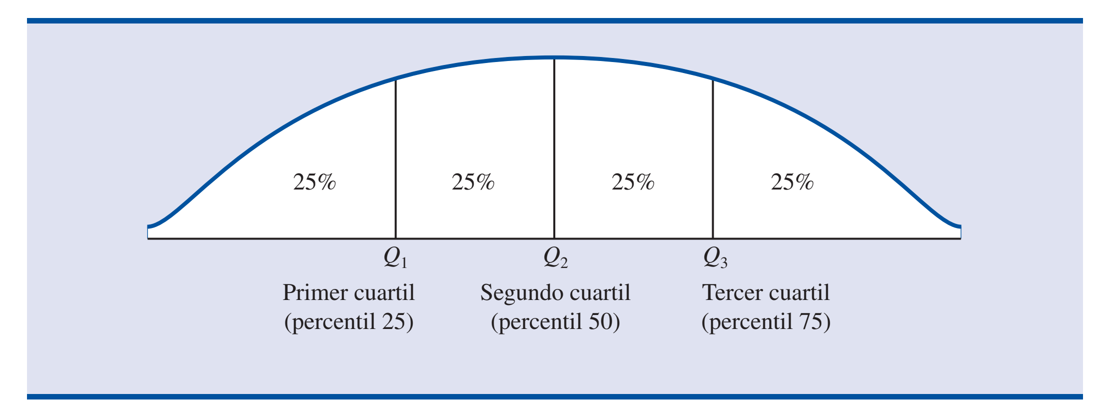
Paso 2.

$$i = \left(\frac{p}{100}\right)n = \left(\frac{85}{100}\right)12 = 10.2$$

Paso 3. Como i no es un número entero, se debe *redondear.* La posición del percentil 85 es el primer entero mayor que 10.2, es la posición 11.

Observe ahora los datos, entonces el percentil 85 es el dato en la posición 11, o sea 3730.

FIGURA 3.1 LOCALIZACIÓN DE LOS CUARTILES



Para ampliar la formación en el uso de este procedimiento, calculará el percentil 50 en los sueldos mensuales iniciales. Al aplicar el paso 2 obtiene.

$$i = \left(\frac{50}{100}\right)12 = 6$$

Como i es un número entero, de acuerdo con el paso 3 b) el percentil 50 es el promedio de los valores de los datos que se encuentran en las posiciones seis y siete; de manera que el percentil 50 es (3490 + 3520)/2 = 3505. Observe que el *percentil 50 coincide con la mediana.*

Cuartiles

Los cuartiles sólo son percentiles determinados; así que los pasos para calcular los percentiles también se emplean para calcular los cuartiles.

Con frecuencia es conveniente dividir los datos en cuatro partes; así, cada parte contiene una cuarta parte o 25% de las observaciones. En la figura 3.1 se muestra una distribución de datos dividida en cuatro partes. A los puntos de división se les conoce como **cuartiles** y están definidos como sigue:

Q_1 = primer cuartil, o percentil 25
Q_2 = segundo cuartil, o percentil 50
Q_3 = tercer cuartil, o percentil 75

Una vez más se ordenan los sueldos iniciales de menor a mayor. Q_2, el segundo cuartil (la mediana), ya se tiene identificado, es 3505.

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Para calcular los cuartiles Q_1 y Q_3 use la regla para hallar el percentil 25 y el percentil 75. A continuación se presentan estos cálculos.

Para hallar Q_1,

$$i = \left(\frac{p}{100}\right)n = \left(\frac{25}{100}\right)12 = 3$$

Como i es un entero, el paso 3 b) indica que el primer cuartil, o el percentil 25, es el promedio del tercer y cuarto valores de los datos; esto es, Q_1 = (3450 + 3480)/2 = 3465.

Para hallar Q_3,

$$i = \left(\frac{p}{100}\right)n = \left(\frac{75}{100}\right)12 = 9$$

Como i es un entero, el paso 3 b) indica que el tercer cuartil, o el percentil 75, es el promedio del noveno y décimo valores de los datos; esto es, Q_3 = (3550 + 3650)/2 = 3600.

Los cuartiles dividen los datos de los sueldos iniciales en cuatro partes y cada parte contiene 25% de las observaciones.

3310 3355 3450 | 3480 3480 3490 | 3520 3540 3550 | 3650 3730 3925

$Q_1 = 3465$ $\quad Q_2 = 3505$ (Mediana) $\quad Q_3 = 3600$

Los cuartiles han sido definidos como el percentil 25, el percentil 50 y el percentil 75. Por lo que los cuartiles se calculan de la misma manera que los percentiles. Sin embargo, algunas veces se siguen otras convenciones para calcular los cuartiles, por ello los valores que se dan para los cuartiles varían ligeramente, dependiendo de la convención que se siga. De cualquier manera, el objetivo de calcular los cuartiles siempre es dividir los datos en cuatro partes iguales.

NOTAS Y COMENTARIOS

Cuando el conjunto de datos contiene valores extremos, es preferible usar la mediana que la media como unidad de localización central. Otra medida que suele ser usada cuando hay valores extremos es la *media recortada*. La media recortada se obtiene eliminando del conjunto de datos un determinado porcentaje de los valores menores y mayores y calculando después la media de los valores restantes. Por ejemplo, la media recortada a 5% se obtiene eliminando el 5% menor y el 5% mayor de los valores y calculando después la media de los valores restantes. Con la muestra de los 12 sueldos iniciales, 0.05(12) = 0.6. Redondear este valor a 1, indica que en la media recortada a 5% se elimina el valor (1) menor y el valor (1) mayor. La media recortada a 5% usando las 10 observaciones restantes es 3524.50.

Ejercicios

Método

1. Los valores de los datos en una muestra son 10, 20, 12, 17 y 16. Calcule la media y la mediana.
2. Los datos en una muestra son 10, 20, 21, 17, 16 y 25. Calcule la media y la mediana.
3. Los valores en una muestra son 27, 25, 20, 15, 30, 34, 28 y 25. Calcule los percentiles 20, 25, 65 y 75

Autoexamen

4. Una muestra tiene los valores 53, 55, 70, 58, 64, 57, 53, 69, 57, 68 y 53. Calcule la media, la mediana y la moda.

Aplicaciones

5. El Dow Jones Travel Index informa sobre lo que pagan por noche en un hotel en las principales ciudades de Estados Unidos los viajeros de negocios (*The Wall Street Journal*, 16 de enero de 2004). Los precios promedio por noche en 20 ciudades son los siguientes:



Hotels

Atlanta	$163	Minneapolis	$125
Boston	177	New Orleans	167
Chicago	166	New York	245
Cleveland	126	Orlando	146
Dallas	123	Phoenix	139
Denver	120	Pittsburgh	134
Detroit	144	San Francisco	167
Houston	173	Seattle	162
Los Angeles	160	St. Louis	145
Miami	192	Washington, D.C.	207

a. ¿Cuál es la media en el precio de estas habitaciones?
b. ¿Cuál es la mediana en el precio de estas habitaciones?
c. ¿Cuál es la moda?
d. ¿Cuál es el primer cuartil?
e. ¿Cuál es el tercer cuartil?

6. Una asociación recaba información sobre sueldos anuales iniciales de los recién egresados de universidades de acuerdo con su especialidad. El salario anual inicial de los administradores de empresas es $39 580 (*CNNMoney.com*, 15 de febrero de 2006). A continuación se presentan muestras de los sueldos anuales iniciales de especialistas en marketing y en contaduría (los datos están en miles):



Egresados de marketing

34.2	45.0	39.5	28.4	37.7	35.8	30.6	35.2	34.2	42.4

Egresados de contaduría

33.5	57.1	49.7	40.2	44.2	45.2	47.8	38.0
53.9	41.1	41.7	40.8	55.5	43.5	49.1	49.9

a. Para cada uno de los grupos de sueldos iniciales calcule moda, mediana y media.
b. Para cada uno de los grupos de sueldos iniciales calcule el primer y el tercer cuartil.
c. Los egresados de contaduría suelen tener mejores salarios iniciales. ¿Qué indican los datos muestrales acerca de la diferencia entre los sueldos anuales iniciales de egresados de marketing y de contaduría?

7. La Asociación Estadounidense de Inversionistas Individuales realiza una investigación anual sobre los corredores de bolsa (*AAII Journal,* enero de 2003). En la tabla 3.2 se muestran las comisiones que cobran los corredores de bolsa con descuento por dos tipos de transacciones: transacción con ayuda del corredor de 100 acciones a $50 por acción y transacción en línea de 500 acciones a $50 por acción.
a. Calcule la media, mediana y moda de las comisiones que se cobran por una transacción con ayuda del corredor de 100 acciones a $50 por acción.
b. Calcule la media, mediana y moda de las comisiones que se cobran por una transacción en línea de 500 acciones a $50 por acción.
c. ¿Qué cuesta más, una transacción con ayuda del corredor de 100 acciones a $50 por acción o una transacción en línea de 500 acciones a $50 por acción?
d. ¿Está relacionado el costo de la transacción con el monto de la transacción?

TABLA 3.2 COMISIONES QUE COBRAN LOS CORREDORES DE BOLSA



Corredor	Con ayuda del corredor de 100 acciones $50/acción	En línea 500 acciones a $50/acción	Corredor	Con ayuda del corredor de 100 acciones $50/acción	En línea 500 acciones a $50/acción
Accutrade	30.00	29.95	Merrill Lynch Direct	50.00	29.95
Ameritrade	24.99	10.99	Muriel Siebert	45.00	14.95
Banc of America	54.00	24.95	NetVest	24.00	14.00
Brown & Co.	17.00	5.00	Recom Securities	35.00	12.95
Charles Schwab	55.00	29.95	Scottrade	17.00	7.00
CyberTrader	12.95	9.95	Sloan Securities	39.95	19.95
E*TRADE Securities	49.95	14.95	Strong Investments	55.00	24.95
First Discount	35.00	19.75	TD Waterhouse	45.00	17.95
Freedom Investments	25.00	15.00	T. Rowe Price	50.00	19.95
Harrisdirect	40.00	20.00	Vanguard	48.00	20.00
Investors National	39.00	62.50	Wall Street Discount	29.95	19.95
MB Trading	9.95	10.55	York Securities	40.00	36.00

Fuente: AAII Journal, enero de 2003.

8. Millones de estadounidenses trabajan para sus empresas desde sus hogares. A continuación se presenta una muestra de datos que dan las edades de estas personas que trabajan desde sus hogares.



18	54	20	46	25	48	53	27	26	37
40	36	42	25	27	33	28	40	45	25

a. Calcule la media y la moda.
b. La edad mediana de la población de todos los adultos es de 36 años (*The World Almanac,* 2006). Use la edad mediana de los datos anteriores para decir si las personas que trabajan desde sus hogares tienden a ser más jóvenes o más viejos que la población de todos los adultos.
c. Calcule el primer y el tercer cuartil.
d. Calcule e interprete el percentil 32.

9. J. D. Powers and Associates hicieron una investigación sobre el número de minutos por mes que los usuarios de teléfonos celulares usan sus teléfonos (Associated Press, junio de 2002). A continuación se muestran los minutos por mes hallados en una muestra de 15 usuarios de teléfonos celulares

615	135	395
430	830	1180
690	250	420
265	245	210
180	380	105

a. ¿Cuál es la media de los minutos de uso por mes?
b. ¿Cuál es la mediana de los minutos de uso por mes?
c. ¿Cuál es el percentil 85?
d. J. D. Powers and Associates informa que los planes promedio para usuarios de celulares permiten hasta 750 minutos de uso por mes. ¿Qué indican los datos acerca de la utilización que hacen los usuarios de teléfonos celulares de sus planes mensuales?

10. En una investigación hecha por la Asociación Estadounidense de Hospitales se encontró que la mayor parte de las salas de emergencias de los hospitales estaban operando a toda su capacidad (Associated Press, 9 de abril de 2002). En esta investigación se reunieron datos de los tiempos de espera en las salas de emergencias de hospitales donde éstas operaban a toda su capacidad y de hospitales en que operan de manera equilibrada y rara vez manejan toda su capacidad.

Tiempos de espera para las SE en hospitales a toda capacidad		**Tiempos de espera para las SE en hospitales en equilibrio**	
87	59	60	39
80	110	54	32
47	83	18	56
73	79	29	26
50	50	45	37
93	66	34	38
72	115		

a. Calcule la media y la mediana de estos tiempos de espera en los hospitales a toda capacidad.
b. Calcule la media y la mediana de estos tiempos de espera en los hospitales en equilibrio.
c. Con base en estos resultados, ¿qué observa acerca de los tiempos de espera para las salas de emergencia? ¿Preocuparán a la Asociación Estadounidense de Hospitales los resultados estadísticos encontrados aquí?

11. En una prueba sobre consumo de gasolina se examinaron a 13 automóviles en un recorrido de 100 millas, tanto en ciudad como en carretera. Se obtuvieron los datos siguientes de rendimiento en millas por galón.

Ciudad:	16.2	16.7	15.9	14.4	13.2	15.3	16.8	16.0	16.1	15.3	15.2	15.3	16.2
Carretera:	19.4	20.6	18.3	18.6	19.2	17.4	17.2	18.6	19.0	21.1	19.4	18.5	18.7

Use la media, la mediana y la moda para indicar cuál es la diferencia en el consumo entre ciudad y carretera.

12. La empresa Walt Disney compró en 7.4 mil millones de dólares Pixar Animation Studios Inc. (*CNNMoney.com* 24 de enero de 2006). A continuación se presentan las películas animadas producidas por cada una de estas empresas (Disney y Pixar). Las ganancias están en millones de dólares. Calcule las ganancias totales, la media, la mediana y los cuartiles para comparar el éxito de las películas producidas por ambas empresas. ¿Sugieren dichos estadísticos por lo menos una razón por la que Disney haya podido estar interesada en comprar Pixar? Analice.



Disney

Películas de Disney	**Ganancias (millones de $)**	**Películas de Pixar**	**Ganancias (millones de $)**
Pocahontas	346	*Toy Story*	362
Hunchback of Notre Dame	325	*A Bug's Life*	363
Hercules	253	*Toy Story 2*	485
Mulan	304	*Monsters, Inc.*	525
Tarzan	448	*Finding Nemo*	865
Dinosaur	354	*The Incredibles*	631
The Emperor's New Groove	169		
Lilo & Stitch	273		
Treasure Planet	110		
The Jungle Book 2	136		
Brother Bear	250		
Home on the Range	104		
Chicken Little	249		

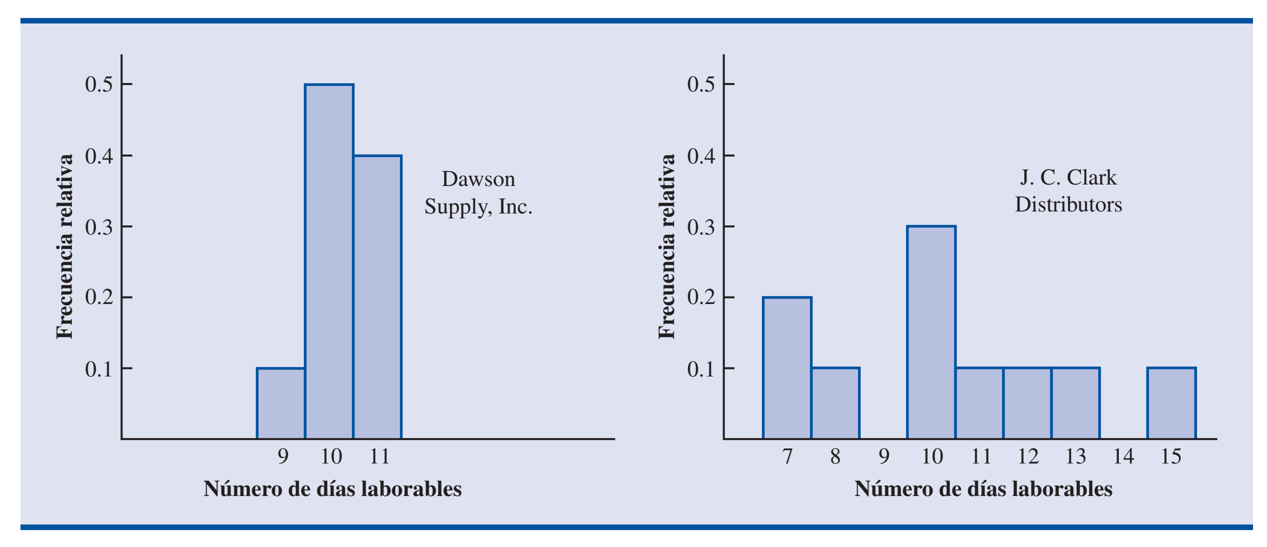
3.2 Medidas de variabilidad

La variabilidad en los tiempos de entrega produce incertidumbre en la planeación de la producción. Los métodos que se presentan en esta sección ayudan a medir y entender la variabilidad.

Además de las medidas de localización, suele ser útil considerar las medidas de variabilidad o de dispersión. Suponga que usted es el encargado de compras de una empresa grande y que con regularidad envía órdenes de compra a dos proveedores. Después de algunos meses de operación, se percata de que el número promedio de días que ambos proveedores requieren para surtir una orden es 10 días. En la figura 3.2 se presentan los histogramas que muestran el número de días que cada uno de los proveedores necesita para surtir una orden. Aunque en ambos casos este número promedio de días es 10 días, ¿muestran los dos proveedores el mismo grado de confiabilidad en términos de tiempos para surtir los productos? Observe la dispersión, o variabilidad, de estos tiempos en ambos histogramas. ¿Qué proveedor preferiría usted?

Para la mayoría de las empresas es importante recibir a tiempo los materiales que necesitan para sus procesos. En el caso de J. C. Clark Distributors sus tiempos de entrega, de siete u ocho días, parecen muy aceptables; sin embargo, sus pocos tiempos de entrega de 13 a 15 días resul-

FIGURA 3.2 DATOS HISTÓRICOS QUE MUESTRAN EL NÚMERO DE DÍAS REQUERIDOS PARA COMPLETAR UNA ORDER



tan desastrosos en términos de mantener ocupada a la fuerza de trabajo y de cumplir con el plan de producción. Este ejemplo ilustra una situación en que la variabilidad en los tiempos de entrega puede ser la consideración más importante en la elección de un proveedor. Para la mayor parte de los encargados de compras, la poca variabilidad que muestra en los tiempos de entrega de Dawson Supply, Inc. hará de esta empresa el proveedor preferido.

Ahora mostramos el estudio de algunas de las medidas de variabilidad más usadas.

Rango

La medida de variabilidad más sencilla es el **rango**.

RANGO

$$\text{Rango} = \text{Valor mayor} - \text{Valor menor}$$

De regreso a los datos de la tabla 3.1 sobre sueldos iniciales de los recién egresados de la carrera de administración, el mayor sueldo inicial es 3925 y el menor 3310. El rango es $3925 - 3310 = 615$.

Aunque el rango es la medida de variabilidad más fácil de calcular, rara vez se usa como única medida. La razón es que el rango se basa sólo en dos observaciones y, por tanto, los valores extremos tienen una gran influencia sobre él. Suponga que uno de los recién egresados haya tenido $10 000 como sueldo inicial, entonces el rango será $10\,000 - 3310 = 6690$ en lugar de 615. Un valor así no sería muy descriptivo de la variabilidad de los datos ya que 11 de los 12 sueldos iniciales se encuentran entre 3310 y 3730.

Rango intercuartílico

Una medida que no es afectada por los valores extremos es el **rango intercuartílico (RIC)**. Esta medida de variabilidad es la diferencia entre el tercer cuartil Q_3 y el primer cuartil Q_1. En otras palabras, el rango intercuartílico es el rango en que se encuentra el 50% central de los datos.

RANGO INTERCUARTÍLICO

$$\text{IQR} = Q_3 - Q_1 \quad \textbf{(3.3)}$$

En los datos de los sueldos mensuales iniciales, los cuartiles son $Q_3 = 3600$ y $Q_1 = 3465$. Por lo tanto el rango intercuartílico es $3600 - 3465 = 135$.

Varianza

La **varianza** es una medida de variabilidad que utiliza todos los datos. La varianza está basada en la diferencia entre el valor de cada observación (x_i) y la media. A la diferencia entre cada valor x_i y la media ($\bar{x}$ cuando se trata de una muestra, μ cuando se trata de una población) se le llama *desviación respecto de la media.* Si se trata de una muestra, una desviación respecto de la media se escribe $(x_i - \bar{x})$, y si se trata de una población se escribe $(x_i - \mu)$. Para calcular la varianza, estas desviaciones respecto de la media *se elevan al cuadrado*.

Si los datos son de una población, el promedio de estas desviaciones elevadas al cuadrado es la *varianza poblacional.* La varianza poblacional se denota con la letra griega σ^2. En una población en la que hay N observaciones y la media poblacional es μ, la varianza poblacional se define como sigue.

VARIANZA POBLACIONAL

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \quad \textbf{(3.4)}$$

En la mayor parte de las aplicaciones de la estadística, los datos a analizar provienen de una muestra. Cuando se calcula la varianza muestral, lo que interesa es estimar la varianza poblacional σ^2. Aunque una explicación detallada está más allá del alcance de este libro, es posible demostrar que si la suma de los cuadrados de las desviaciones respecto de la media se divide entre $n - 1$, en lugar de entre n, la varianza muestral que se obtiene constituye un estimador no sesgado de la varianza poblacional. Por esta razón, la *varianza muestral*, que se denota por s^2, se define como sigue.

La varianza muestral s^2 es el estimador de la varianza poblacional σ^2 .

VARIANZA MUESTRAL

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \quad \textbf{(3.5)}$$

Para ilustrar el cálculo de la varianza muestral, se emplean los datos de los tamaños de cinco grupos de una universidad, presentados en la sección 3.1. En la tabla 3.3 aparece un resumen de los datos con el cálculo de las desviaciones respecto de la media y de los cuadrados de las desviaciones respecto de la media. La suma de los cuadrados de las desviaciones respecto de la media es $\Sigma(x_i - \bar{x})^2 = 256$. Por tanto, siendo $n - 1 = 4$, la varianza muestral es

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{256}{4} = 64$$

Antes de continuar, hay que hacer notar que las unidades correspondientes a la varianza muestral suelen causar confusión. Como los valores que se suman para calcular la varianza, $(x_i - \bar{x})^2$, están elevados al cuadrado, las unidades correspondientes a la varianza muestral tam-

TABLA 3.3 CÁLCULO DE LAS DESVIACIONES Y DE LOS CUADRADOS DE LAS DESVIACIONES RESPECTO DE LA MEDIA EMPLEANDO LOS DATOS DE LOS TAMAÑOS DE CINCO GRUPOS DE ESTADOUNIDENSES

Número de estudiantes en un grupo (x_i)	Número promedio de alumnos en un grupo ($\bar{x}$)	Desviación respecto a la media ($x_i - \bar{x}$)	Cuadrado de la desviación respecto de la media $(x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

La varianza sirve para comparar la variabilidad de dos o más variables.

bién están *elevadas al cuadrado*. Por ejemplo, la varianza muestral en los datos de la cantidad de alumnos en los grupos es $s^2 = 64$ (estudiantes)^2. Las unidades al cuadrado de la varianza dificultan la comprensión e interpretación intuitiva de los valores numéricos de la varianza. Aquí lo recomendable es entender la varianza como una medida útil para comparar la variabilidad de dos o más variables. Al comparar variables, la que tiene la varianza mayor, muestra más variabilidad. Otra interpretación del valor de la varianza suele ser innecesaria.

Para tener otra ilustración del cálculo de la varianza muestral, considere los sueldos iniciales de 12 recién egresados de la carrera de administración, presentados en la tabla 3.1. En la sección 3.1 se vio que la media muestral de los sueldos mensuales iniciales era 3540. En la tabla 3.4 se muestra el cálculo de la varianza muestral ($s^2 = 27\,440.91$).

TABLA 3.4 CÁLCULO DE LA VARIANZA MUESTRAL CON LOS DATOS DE LOS SUELDOS INICIALES

Sueldo mensual (x_i)	Media muestral ($\bar{x}$)	Desviación respecto de la media ($x_i - \bar{x}$)	Cuadrado de la desviación respecto de la media $(x_i - \bar{x})^2$
3450	3540	−90	8 100
3550	3540	10	100
3650	3540	110	12 100
3480	3540	−60	3 600
3355	3540	−185	34 225
3310	3540	−230	52 900
3490	3540	−50	2 500
3730	3540	190	36 100
3540	3540	0	0
3925	3540	385	148 225
3520	3540	−20	400
3480	3540	−60	3 600
		0	301 850
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

Empleando la ecuación (3.5),

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{301\,850}{11} = 27\,440.91$$

En las tablas 3.3 y 3.4 se presenta la suma, tanto de las desviaciones respecto de la media como de los cuadrados de las desviaciones respecto de la media. En todo conjunto de datos, la suma de las desviaciones respecto de la media será *siempre igual a cero.* Observe que en las tablas 3.3 y 3.4 $\Sigma(x_i - \bar{x}) = 0$. Las desviaciones positivas y las desviaciones negativas se anulan mutuamente haciendo que la suma de las desviaciones respecto a la media sea igual a cero.

Desviación estándar

La **desviación estándar** se define como la raíz cuadrada positiva de la varianza. Continuando con la notación adoptada para la varianza muestral y para la varianza poblacional, se emplea s para denotar la desviación estándar muestral y σ para denotar la desviación estándar poblacional. La desviación estándar se obtiene de la varianza como sigue.

La desviación estándar muestral s es el estimador de la desviación estándar poblacional σ.

DESVIACIÓN ESTÁNDAR

$$\text{Desviación estándar muestral} = s = \sqrt{s^2} \quad \textbf{(3.6)}$$

$$\text{Desviación estándar poblacional} = \sigma = \sqrt{\sigma^2} \quad \textbf{(3.7)}$$

Recuerde que la varianza muestral para los tamaños de cinco grupos de una universidad es $s^2 = 64$. Por tanto, la desviación estándar muestral es $s = \sqrt{64} = 8$. En los datos de los sueldos iniciales, la desviación estándar es $s = \sqrt{27\,440.91} = 165.65$.

La desviación estándar es más fácil de interpretar que la varianza debido a que la desviación estándar se mide en las mismas unidades que los datos.

¿Qué se gana con convertir la varianza en la correspondiente desviación estándar? Recuerde que en la varianza las unidades están elevadas al cuadrado. Por ejemplo, la varianza muestral de los datos de los sueldos iniciales de los egresados de administración es $s^2 = 27{,}440.91$ (dólares)^2. Como la desviación estándar es la raíz cuadrada de la varianza, las unidades de la varianza, dólares al cuadrado, se convierten en dólares en la desviación estándar. Por tanto, la desviación estándar de los sueldos iniciales es $165.65. En otras palabras, la desviación estándar se mide en las mismas unidades que los datos originales. Por esta razón es más fácil comparar la desviación estándar con la media y con otros estadísticos que se miden en las mismas unidades que los datos originales.

Coeficiente de variación

El coeficiente de variación es una medida relativa de la variabilidad; mide la desviación estándar en relación con la media.

En algunas ocasiones se requiere un estadístico descriptivo que indique cuán grande es la desviación estándar en relación con la media. Esta medida es el **coeficiente de variación** y se representa como porcentaje.

COEFICIENTE DE VARIACIÓN

$$\left(\frac{\text{Desviación estándar}}{\text{Media}} \times 100\right)\% \quad \textbf{(3.8)}$$

En los datos de los tamaños de los cinco grupos de estudiantes, se encontró una media muestral de 44 y una desviación estándar muestral de 8. El coeficiente de variación es $[(8/44) \times 100]\% = 18.2\%$. Expresado en palabras, el coeficiente de variación indica que la desviación estándar muestral es 18.2% del valor de la media muestral. En los datos de los sueldos iniciales, la media muestral encontrada es 3540 y la desviación estándar muestral es 165.65, el coeficiente de variación, $[(165.65/3540) \times 100]\% = 4.7\%$, indica que la desviación estándar muestral es sólo 4.7% del valor de la media muestral. En general, el coeficiente de variación es un estadístico útil para comparar la variabilidad de variables que tienen desviaciones estándar distintas y medias distintas.

NOTAS Y COMENTARIOS

1. Los paquetes de software para estadística y las hojas de cálculo sirven para buscar los estadísticos descriptivos presentados en este capítulo. Una vez que los datos se han ingresado en una hoja de cálculo, basta emplear unos cuantos comandos sencillos para obtener los estadísticos deseados. En los apéndices 3.1 y 3.2 se muestra cómo usar Minitab y Excel para lograrlo.
2. La desviación estándar suele usarse como medida del riesgo relacionado con una inversión en acciones o en fondos de acciones (*Bussines-Week,* 7 de enero de 2000). Proporciona una medida de cómo fluctúa la rentabilidad mensual respecto de la rentabilidad promedio a largo plazo.
3. Redondear los valores de la media muestral $\bar{x}$ y de los cuadrados de las desviaciones $(x_i - \bar{x})^2$ puede introducir errores cuando se emplea una calculadora para el cálculo de la varianza y de la desviación estándar. Para reducir los errores de redondeo se recomienda conservar por lo menos seis dígitos significativos en los cálculos intermedios. La varianza o la desviación estándar obtenidos se redondean entonces a menos dígitos significativos.
4. Otra fórmula alterna para el cálculo de la varianza muestral es

$$s^2 = \frac{\Sigma x_i^2 - n\bar{x}^2}{n - 1}$$

donde $\Sigma x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$.

Ejercicios

Métodos

13. Considere una muestra con los datos 10, 20, 12, 17 y 16. Calcule el rango y el rango intercuartílico.

14. Considere una muestra que tiene como valores 10, 20, 12, 17 y 16. Calcule la varianza y la desviación estándar.



15. Considere una muestra con valores 27, 25, 0, 15, 30, 34, 28 y 25. Calcule el rango, el rango intercuartílico, la varianza y la desviación estándar.

Aplicaciones



16. Las puntuaciones obtenidas por un jugador de boliche en seis juegos fueron 182, 168, 184, 190, 170 y 174. Use estos datos como una muestra y calcule los estadísticos descriptivos siguientes
    a. Rango
    b. Varianza
    c. Desviación estándar
    d. Coeficiente de variación

17. A *home theater in a box* es la manera más sencilla y económica de tener sonido envolvente en un centro de entretenimiento en casa. A continuación se presenta una muestra de precios (*Consumer Report Buying Guide* 2004). Los precios corresponden a modelos con y sin reproductor de DVD.

Modelos con reproductor de DVD	Precio	Modelos sin reproductor de DVD	Precio
Sony HT-1800DP	$450	Pioneer HTP-230	$300
Pioneer HTD-330DV	300	Sony HT-DDW750	300
Sony HT-C800DP	400	Kenwood HTB-306	360
Panasonic SC-HT900	500	RCA RT-2600	290
Panasonic SC-MTI	400	Kenwood HTB-206	300

a. Calcule el precio medio de los modelos con reproductor de DVD y el precio medio de los modelos sin reproductor de DVD. ¿Cuánto es lo que se paga de más por tener un reproductor de DVD en casa?
b. Calcule el rango, la varianza y la desviación estándar de las dos muestras. ¿Qué le dice esta información acerca de los precios de los modelos con y sin reproductor de DVD?

18. Las tarifas de renta de automóviles por día en siete ciudades del este de Estados Unidos son las siguientes (*The Wall Street Journal* 16 de enero de 2004).

Ciudad	Tarifa por día
Boston	$43
Atlanta	35
Miami	34
New York	58
Orlando	30
Pittsburgh	30
Washington, D.C.	36

a. Calcule la media, la varianza y la desviación estándar de estas tarifas.
b. En una muestra similar de siete ciudades del oeste la media muestral de las tarifas fue de $38 por día. La varianza y la desviación estándar fueron 12.3 y 3.5 cada una. Analice la diferencia entre las tarifas de las ciudades del este y del oeste.

19. *Los Angeles Times* informa con regularidad sobre el índice de la calidad del aire en varias regiones del sur de California. En una muestra de los índices de calidad del aire en Pomona se tienen los datos siguientes: 28, 42, 58, 48, 45, 55, 60, 49 y 50.
a. Calcule el rango y el rango intercuartílico.
b. Calcule la varianza muestral y la desviación estándar muestral.
c. En una muestra de índices de calidad del aire en Anaheim, la media muestral es 48.5, la varianza muestral es 136 y la desviación estándar muestral es 11.66. Con base en estos estadísticos descriptivos compare la calidad del aire en Pomona y en Anaheim.

20. A continuación se presentan los datos que se usaron para elaborar los histogramas sobre el número de días necesarios para surtir una orden (véase la figura 3.2).

Días de entrega de Dawson Supply, Inc.:	11	10	9	10	11	11	10	11	10	10
Días de entrega de Clark Distributors:	8	10	13	7	10	11	10	7	15	12

Use el rango y la desviación estándar para sustentar la observación hecha antes de que Dawson Supply proporcione los tiempos de entrega más consistentes.

21. ¿Cómo están los costos de abarrotes en el país? A partir de una canasta alimenticia de 10 artículos entre los que se encuentran carne, leche, pan, huevos, café, papas, cereal y jugo de naranja, la revista *Where to Retire* calculó el costo de la canasta alimenticia en seis ciudades y en seis zonas con personas jubiladas en todo el país (*Where to Retire* noviembre/diciembre de 2003). Los datos encontrados, al dólar más cercano, se presentan a continuación.

Ciudad	Costo	Zona de jubilados	Costo
Buffalo, NY	$33	Biloxi-Gulfport, MS	$29
Des Moines, IA	27	Asheville, NC	32
Hartford, CT	32	Flagstaff, AZ	32
Los Angeles, CA	38	Hilton Head, SC	34
Miami, FL	36	Fort Myers, FL	34
Pittsburgh, PA	32	Santa Fe, NM	31

a. Calcule la media, varianza y desviación estándar de las ciudades y de las zonas de jubilados.
b. ¿Qué observaciones puede hacer con base en estas dos muestras?



Broker

22. La Asociación Estadounidense de Inversionistas Individuales realiza cada año una investigación sobre los corredores de bolsa con descuento (*AAII Journal*, enero de 2003). En la tabla 3.2 se muestran las comisiones que cobran 24 corredores de bolsa con descuento por dos tipos de transacciones: transacción con ayuda del corredor de 100 acciones a $50 la acción y transacción en línea de 500 acciones a $50 la acción.
    a. Calcule el rango y el rango intercuartílico en cada tipo de transacción.
    b. Calcule la varianza y la desviación estándar en cada tipo de transacción.
    c. Calcule el coeficiente de variación en cada tipo de transacción.
    d. Compare la variabilidad en el costo que hay en los dos tipos de transacciones
24. Las puntuaciones de un jugador de golf en el 2005 y 2006 son las siguientes:

2005	74	78	79	77	75	73	75	77
2006	71	70	75	77	85	80	71	79

    a. Use la media y la desviación estándar para evaluar a este jugador de golf en estos dos años.
    b. ¿Cuál es la principal diferencia en su desempeño en estos dos años? ¿Se puede ver algún progreso en sus puntuaciones del 2006?, ¿cuál?
24. Los siguientes son los tiempos que hicieron los velocistas de los equipos de pista y campo de una universidad en un cuarto de milla y en una milla (los tiempos están en minutos).

Tiempos en un cuarto de milla:	0.92	0.98	1.04	0.90	0.99
Tiempos en una milla:	4.52	4.35	4.60	4.70	4.50

Después de ver estos datos, el entrenador comentó que en un cuarto de milla los tiempos eran más homogéneos. Use la desviación estándar y el coeficiente de variación para resumir la variabilidad en los datos. El uso del coeficiente de variación, ¿indica que la aseveración del entrenador es correcta?



Medidas de la forma de la distribución, de la posición relativa y de la detección de observaciones atípicas

Se han descrito ya varias medidas de localización y de variabilidad de los datos. Además de estas medidas se necesita una medida de la forma de la distribución. En el capítulo 2 se vio que un histograma es una representación gráfica que muestra la forma de una distribución. Una medida numérica importante de la forma de una distribución es el **sesgo**.

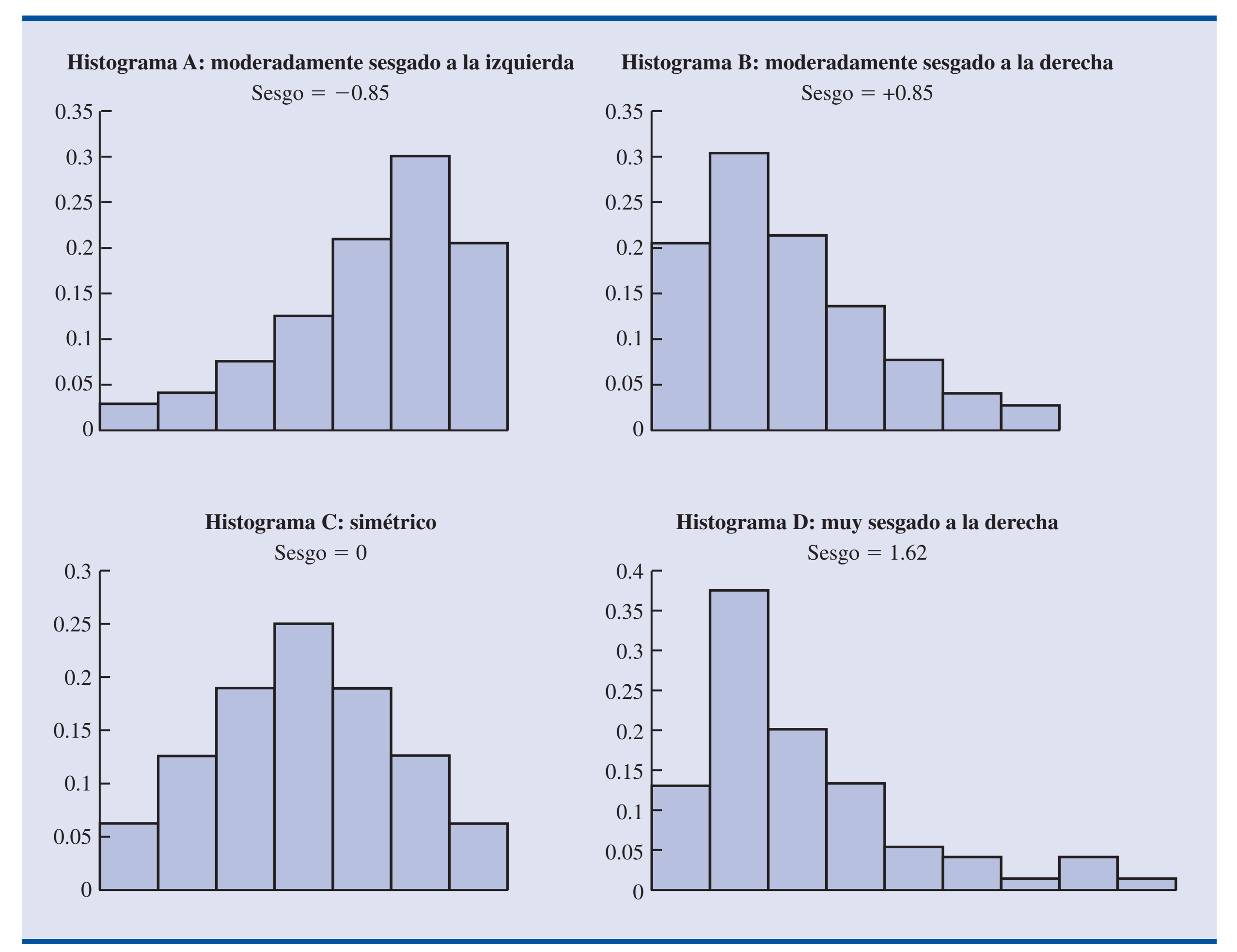
Forma de la distribución

En la figura 3.3 se muestran cuatro histogramas elaborados a partir de distribuciones de frecuencias relativas. Los histogramas A y B son moderadamente sesgados. El histograma A es sesgado a la izquierda, su sesgo es −0.85. El histograma B es sesgado a la derecha, su sesgo es +0.85. El histograma C es simétrico; su sesgo es cero. El histograma D es muy sesgado a la derecha; su sesgo es 1.62. La fórmula que se usa para calcular el sesgo es un poco complicada.* Sin embargo, es fácil de calcular empleando el software para estadística (véase los apéndices 3.1 y 3.2). En

*La fórmula para calcular el sesgo de datos muestrales es:

$$\text{Sesgo} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

FIGURA 3.3 HISTOGRAMAS QUE MUESTRAN EL SESGO DE CUATRO DISTRIBUCIONES



los datos sesgados a la izquierda, el sesgo es negativo; en datos sesgados a la derecha, el sesgo es positivo. Si los datos son simétricos, el sesgo es cero.

En una distribución simétrica, la media y la mediana son iguales. Si los datos están sesgados a la derecha, la media será mayor que la mediana; si los datos están sesgados a la izquierda, la media será menor que la mediana. Los datos que se emplearon para elaborar el histograma D son los datos de las compras realizadas en una tienda de ropa para dama. El monto medio de las compras es $77.60 y el monto mediano de las compras es $59.70. Los pocos montos altos de compras tienden a incrementar la media, mientras que a la mediana no le afectan estos montos elevados de compras. Cuando los datos están ligeramente sesgados, se prefiere la mediana como medida de localización.

Puntos *z*

Además de las medidas de localización, variabilidad y forma, interesa conocer también la ubicación relativa de los valores de un conjunto de datos. Las medidas de localización relativa ayudan a determinar qué tan lejos de la media se encuentra un determinado valor.

A partir de la media y la desviación estándar, se puede determinar la localización relativa de cualquier observación. Suponga que tiene una muestra de n observaciones, en que los valores se

denotan $x_1, x_2, \ldots, x_n$. Suponga además que ya determinó la media muestral, que es $\bar{x}$ y la desviación estándar muestral, que es s. Para cada valor x_i existe otro valor llamado **punto z**. La ecuación (3.9) permite calcular el punto z correspondiente a cada x_i.

PUNTO z

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.9}$$

donde

z_i = punto z para x_i
$\bar{x}$ = media muestral
s = desviación estándar muestral

Al punto z también se le suele llamar *valor estandarizado*. El punto z_i puede ser interpretado como el *número de desviaciones estándar a las que x_i se encuentra de la media $\bar{x}$*. Por ejemplo si $z_1 = 1.2$, esto indica que x_1 es 1.2 desviaciones estándar mayor que la media muestral. De manera similar, $z_2 = -0.5$ indica que x_2 es 0.5 o 1/2 desviación estándar menor que la media muestral. Puntos z mayores a cero corresponden a observaciones cuyo valor es mayor a la media, y puntos z menores que cero corresponden a observaciones cuyo valor es menor a la media. Si el punto z es cero, el valor de la observación correspondiente es igual a la media.

El punto z de cualquier observación se interpreta como una medida relativa de la localización de la observación en el conjunto de datos. Por tanto, observaciones de dos conjuntos de datos distintos que tengan el mismo punto z tienen la misma localización relativa; es decir, se encuentran al mismo número de desviaciones estándar de la media.

En la tabla 3.5 se calculan los puntos z correspondientes a los tamaños de los grupos de estudiantes. Recuerde que ya calculó la media muestral, $\bar{x} = 44$, y la desviación estándar muestral, $s = 8$. El punto z de la quinta observación, que es –1.50, indica que esta observación está más alejada de la media; esta observación está 1.50 desviaciones estándar más abajo de la media.

Teorema de Chebyshev

El **teorema de Chebyshev** permite decir qué proporción de los valores que se tienen en los datos debe estar dentro de un determinado número de desviaciones estándar de la media.

TABLA 3.5 PUNTOS z CORRESPONDIENTES A LOS DATOS DE LOS TAMAÑOS DE LOS GRUPOS DE ESTUDIANTES

Número de estudiantes en un grupo (x_i)	Desviación respecto de la media ($x_i - \bar{x}$)	Puntos z $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	−2	−2/8 = −0.25
46	2	2/8 = 0.25
32	−12	−12/8 = −1.50

TEOREMA DE CHEBYSHEV

Por lo menos $(1 - 1/z^2)$ de los valores que se tienen en los datos deben encontrarse dentro de z desviaciones estándar de la media, donde z es cualquier valor mayor que 1.

De acuerdo con este teorema para $z = 2$, 3 y 4 desviaciones estándar se tiene

- Por lo menos 0.75, o 75%, de los valores de los datos deben estar dentro de $z = 2$ desviaciones estándar de la media.
- Al menos 0.89, o 89%, de los valores deben estar dentro de $z = 3$ desviaciones estándar de la media.
- Por lo menos 0.94, o 94%, de los valores deben estar dentro de $z = 4$ desviaciones estándar de la media.

Para dar un ejemplo del uso del teorema de Chebyshev, suponga que en las calificaciones obtenidas por 100 estudiantes en un examen de estadística para la administración, la media es 70 y la desviación estándar es 5. ¿Cuántos estudiantes obtuvieron puntuaciones entre 60 y 80?, ¿y cuántos tuvieron puntuaciones entre 58 y 82?

En el caso de las puntuaciones entre 60 y 80 observe que 60 está dos desviaciones estándar debajo de la media y que 80 está dos desviaciones estándar sobre la media. Mediante el teorema de Chebyshev encuentre que por lo menos 0.75, o por lo menos 75%, de las observaciones deben tener valores dentro de dos desviaciones estándar de la media. Así que por lo menos 75% de los estudiantes deben haber tenido puntuaciones entre 60 y 80.

En el teorema de Chebyshev se requiere que $z > 1$, pero z no tiene que ser entero.

En el caso de las puntuaciones entre 58 y 82, se encuentra que $(58 - 70)/5 = -2.4$, por lo que 58 se encuentra 2.4 desviaciones estándar debajo de la media, y que $(82 - 70)/5 = +2.4$, entonces 82 se encuentra 2.4 desviaciones estándar sobre la media. Al aplicar el teorema de Chebyshev con $z = 2.4$, se tiene

$$\left(1 - \frac{1}{z^2}\right) = \left(1 - \frac{1}{(2.4)^2}\right) = 0.826$$

Por lo menos 82.6% de los estudiantes deben tener puntuaciones entre 58 y 82.

Regla empírica

La regla empírica está basada en la distribución de probabilidad normal, la cual se estudiará en el capítulo 6. La distribución normal se emplea mucho en todo el libro

Una de las ventajas del teorema de Chebyshev es que se aplica a cualquier conjunto de datos, sin importar la forma de la distribución de los datos. En efecto se usa para cualquiera de las distribuciones de la figura 3.3. Sin embargo, en muchas aplicaciones prácticas los datos muestran una distribución simétrica con forma de montaña o de campana como en la figura 3.4. Cuando se cree que los datos tienen aproximadamente esta distribución, se puede emplear la **regla empírica** para determinar el porcentaje de los valores de los datos que deben encontrarse dentro de un determinado número de desviaciones estándar de la media.

REGLA EMPÍRICA

Cuando los datos tienen una distribución en forma de campana:

- Cerca de 68% de los valores de los datos se encontrarán a no más de una desviación estándar desde la media.
- Aproximadamente 95% de los valores de los datos se encontrarán a no más de dos desviaciones estándar desde la media.
- Casi todos los valores de los datos estarán a no más de tres desviaciones estándar de la media.

FIGURA 3.4 DISTRIBUCIÓN EN FORMA DE MONTAÑA O DE CAMPANA

Por ejemplo, los envases con detergente líquido se llenan en forma automática en una línea de producción. Los pesos de llenado suelen tener una distribución en forma de campana. Si el peso medio de llenado es de 16 onzas y la desviación estándar de 0.25 onzas, la regla empírica es aplicada para sacar las conclusiones siguientes:

- Aproximadamente 68% de los envases llenados pesarán entre 15.75 y 16.25 onzas (estarán a no más de una desviación estándar de la media).
- Cerca de 95% de los envases llenados pesarán entre 15.50 y 16.50 onzas (estarán a no más de dos desviaciones estándar de la media).
- Casi todos los envases llenados pesarán entre 15.25 y 16.75 onzas (estarán a no más de tres desviaciones estándar de la media).

Detección de observaciones atípicas

Algunas veces un conjunto de datos tiene una o más observaciones cuyos valores son mucho más grandes o mucho más pequeños que la mayoría de los datos. A estos valores extremos se les llama **observaciones atípicas**. Las personas que se dedican a la estadística y con experiencia en ella toman medidas para identificar estas observaciones atípicas y después las revisan con cuidado. Una observación extraña quizá sea el valor de un dato que se anotó de modo incorrecto. Si es así puede corregirse antes de continuar con el análisis. Una observación atípica tal vez provenga, también, de una observación que se incluyó indebidamente en el conjunto de datos; si es así se puede eliminar. Por último, una observación atípica quizá es un dato con un valor inusual, anotado correctamente y que sí pertenece al conjunto de datos. En tal caso debe conservarse.

Es conveniente determinar si hay observaciones atípicas antes de tomar decisiones con base en el análisis de los datos. Al escribir los datos o al ingresarlos en la computadora suelen cometerse errores. Las observaciones atípicas no necesariamente deben ser eliminadas, pero sí debe verificarse su exactitud y que sean adecuadas.

Para identificar las observaciones atípicas se emplean los valores estandarizados (puntos z). Recuerde que la regla empírica permite concluir que en los datos con una distribución en forma de campana, casi todos los valores se encuentran a no más de tres desviaciones estándar de la media. Por tanto, si usa los puntos z para identificar las observaciones atípicas, es recomendable considerar cualquier dato cuyo punto z sea menor que -3 o mayor que $+3$ como una observación atípica. Debe examinar la exactitud de tales valores y si en realidad pertenecen al conjunto de datos.

De regreso a los puntos z correspondientes a los datos de los tamaños de grupos de estudiantes de la tabla 3.5, la puntuación -1.50 indica que el tamaño del quinto grupo es el que se encuentra más alejado de la media. Sin embargo, este valor estandarizado queda completamente dentro de los límites de -3 y $+3$. Por tanto, los puntos z no indican que haya observaciones atípicas en estos datos.

NOTAS Y COMENTARIOS

1. El teorema de Chebyshev es aplicable a cualquier conjunto de datos y se usa para determinar el número mínimo de los valores de los datos que estarán a no más de un determinado nú-

mero de desviaciones estándar de la media. Si se sabe que los datos tienen forma de campana se puede decir más. Por ejemplo, la regla empírica permite decir que *cerca de* 95% de los valores de los datos estarán a no más de dos desviaciones estándar de la media. El teorema de Chebyshev sólo permite concluir que por lo menos 75% de los valores de los datos estarán en ese intervalo.

2. Antes de analizar un conjunto de datos, los estadísticos suelen hacer diversas verificaciones para confirmar la validez de los datos. En estudios grandes no es poco común que se cometan errores al anotar los datos o al ingresarlos en la computadora. Identificar las observaciones atípicas es una herramienta usada para verificar la validez de los datos.

Ejercicios

Métodos

25. Considere una muestra cuyos datos tienen los valores 10, 20, 12, 17 y 16. Calcule el punto z de cada una de estas cinco observaciones.

26. Piense en una muestra en que la media es 500 y la desviación estándar es 100. ¿Cuáles son los puntos z de los datos siguientes: 520, 650, 500, 450 y 280?

Autoexamen

27. Considere una muestra en que la media es 30 y la desviación estándar es 5. Utilice el teorema de Chebyshev para determinar el porcentaje de los datos que se encuentra dentro de cada uno de los rangos siguientes.
    a. 20 a 40
    b. 15 a 45
    c. 22 a 38
    d. 18 a 42
    e. 12 a 48

28. Suponga datos que tienen una distribución en forma de campana cuya media es 30 y desviación estándar 5. Utilice la regla empírica para determinar el porcentaje de los datos que se encuentra dentro de cada uno de los rangos siguientes.
    a. 20 a 40
    b. 15 a 45
    c. 25 a 35

Aplicaciones



29. En una encuesta nacional se encontró que los adultos duermen en promedio 6.9 horas por noche. Suponga que la desviación estándar es 1.2 horas.
    a. Emplee el teorema de Chebyshev para hallar el porcentaje de individuos que duermen entre 4.5 y 9.3 horas.
    b. Mediante el teorema de Chebyshev encuentre el porcentaje de individuos que duermen entre 3.9 y 9.9 horas.
    c. Suponga que el número de horas de sueño tiene una distribución en forma de campana. Use la regla empírica para calcular el porcentaje de individuos que duermen entre 4.5 y 9.3 horas por día. Compare este resultado con el valor que obtuvo en el inciso a empleando este resultado.

30. La Administración de Información de Energía informó que el precio medio del galón de gasolina fue $2.30 (*Energy Information Administration,* 27 de febrero de 2006). Admita que la desviación estándar haya sido $0.10 y que el precio del galón de gasolina tenga una distribución en forma de campana.
    a. ¿Qué porcentaje de la gasolina se vendió entre $2.20 y $2.40 por galón?
    b. ¿Qué porcentaje de la gasolina se vendió entre $2.20 y $2.50 por galón?
    c. ¿Qué porcentaje de la gasolina se vendió a más de $2.50 por galón?

31. El promedio de los puntos obtenidos en una sección de un examen a nivel nacional fue 507. Si la desviación estándar es aproximadamente 100, conteste las preguntas siguientes usando una distribución en forma de campana y la regla empírica.

a. ¿Qué porcentaje de los estudiantes obtuvo una puntuación superior a 607?
b. ¿Qué porcentaje de los estudiantes obtuvo una puntuación superior a 707?
c. ¿Qué porcentaje de los estudiantes obtuvo una puntuación entre 407 y 507?
d. ¿Qué porcentaje de los estudiantes obtuvo una puntuación entre 307 y 607?

32. En California los altos costos del mercado inmobiliario han obligado a las familias que no pueden darse el lujo de comprar casas grandes, a construir cobertizos como extensión alternativa de sus viviendas. Estos cobertizos suelen aprovecharse como oficinas, estudios de arte, áreas recreativas, etc. El precio medio de un cobertizo es de $3100 (*Newsweek*, 29 de septiembre de 2003). Asuma que la desviación estándar es de $1200.
a. ¿Cuál es el punto z de un cobertizo cuyo precio es de $2300?
b. ¿Cuál es el punto z de un cobertizo cuyo precio es de $4900?
c. Interprete los valores z de los incisos a y b. Diga si alguno de ellos debe ser considerado como una observación atípica.
d. El artículo de *Newsweek* describe una combinación oficina-cobertizo cuyo precio fue de $13 000. ¿Puede considerar este precio como una observación atípica? Explique.

33. La empresa de luz y fuerza de Florida tiene fama de que después de las tormentas repara muy rápidamente sus líneas. Sin embargo en la época de huracanes del 2004 y 2005, la realidad fue otra, su rapidez para reparar sus líneas no fue suficientemente buena (*The Wall Street Journal*, 16 de enero de 2006). Los siguientes datos son de los días que fueron necesarios para restablecer el servicio después de los huracanes del 2004 y 2005.

Huracán	**Días para restablecer el servicio**
Charley	13
Frances	12
Jeanne	8
Dennis	3
Katrina	8
Rita	2
Wilma	18

Con base en esta muestra de siete, calcule los estadísticos descriptivos siguientes
a. Media, mediana y moda.
b. Rango y desviación estándar.
c. ¿En el caso del huracán Vilma considera el tiempo requerido para restablecer el servicio como una observación atípica?
d. Estos siete huracanes ocasionaron 10 millones de interrupciones del servicio a los clientes. ¿Indican dichas estadísticas que la empresa debe mejorar su servicio de reparación en emergencias? Discuta.

34. A continuación se presentan los puntos que obtuvieron los equipos en una muestra de 10 juegos universitarios de la NCAA (*USA Today*, 26 de febrero de 2004).



NCAA

Equipo ganador	**Puntos**	**Equipo perdedor**	**Puntos**	**Margen de ganancia**
Arizona	90	Oregon	66	24
Duke	85	Georgetown	66	19
Florida State	75	Wake Forest	70	5
Kansas	78	Colorado	57	21
Kentucky	71	Notre Dame	63	8
Louisville	65	Tennessee	62	3
Oklahoma State	72	Texas	66	6

Equipo ganador	Puntos	Equipo perdedor	Puntos	Margen de ganancia
Purdue	76	Michigan State	70	6
Stanford	77	Southern Cal	67	10
Wisconsin	76	Illinois	56	20

a. Calcule la media y la desviación estándar de los puntos obtenidos por los equipos ganadores.
b. Suponga que los puntos obtenidos por los equipos ganadores de la NCAA tienen una distribución en forma de campana. Mediante la media y la desviación estándar halladas en el inciso a, estime cuál es el porcentaje de todos los juegos de la NCAA en que el equipo ganador obtuvo 84 puntos o más. Calcule el porcentaje en todos los juegos de la NCAA en que el equipo ganador obtuvo más de 90 puntos.
c. Aproxime la media y la desviación estándar del margen de ganancia. ¿Hay en estos datos alguna observación atípica? Explique.

35. *Consumer Review* publica en Internet estudios y evaluaciones de diversos productos. La siguiente es una lista de 20 sistemas de sonido con sus evaluaciones (www.audioreview.com). La escala de evaluación es de 1 a 5, siendo 5 lo mejor.



Sistema de sonido	Evaluación	Sistema de sonido	Evaluación
Infinity Kappa 6.1	4.00	ACI Sapphire III	4.67
Allison One	4.12	Bose 501 Series	2.14
Cambridge Ensemble II	3.82	DCM KX-212	4.09
Dynaudio Contour 1.3	4.00	Eosone RSF1000	4.17
Hsu Rsch. HRSW12V	4.56	Joseph Audio RM7si	4.88
Legacy Audio Focus	4.32	Martin Logan Aerius	4.26
Mission 73li	4.33	Omni Audio SA 12.3	2.32
PSB 400i	4.50	Polk Audio RT12	4.50
Snell Acoustics D IV	4.64	Sunfire True Subwoofer	4.17
Thiel CS1.5	4.20	Yamaha NS-A636	2.17

a. Calcule la media y la mediana.
b. Aproxime el primer y el tercer cuartil.
c. Estime la desviación estándar.
d. El sesgo de estos datos es -1.67. Comente la forma de esta distribución.
e. Calcule los puntos z correspondientes a Allison One y a Ommi Audio
f. ¿Hay en estos datos alguna observación atípica? Explique.



Análisis exploratorio de datos

En el capítulo 2 se introdujeron el diagrama de tallo y hojas como una técnica para el análisis exploratorio de datos. Recuerde que el análisis exploratorio de datos permite usar operaciones aritméticas sencillas y representaciones gráficas fáciles de dibujar para resumir datos. En esta sección, para continuar con el análisis exploratorio de datos, se considerarán los resúmenes de cinco números y los diagramas de caja.

Resumen de cinco números

En el **resumen de cinco números** se usan los cinco números siguientes para resumir los datos.

1. El valor menor.
2. El primer cuartil (Q_1).
3. La mediana (Q_2).

4. El tercer cuartil (Q_3).
5. El valor mayor.

La manera más fácil de elaborar un resumen de cinco números es, primero, colocar los datos en orden ascendente. Hecho esto, es fácil identificar el valor menor, los tres cuartiles y el valor mayor. A continuación se presentan los salarios iniciales de los 12 recién egresados de la carrera de administración, que se presentaron en la tabla 3.1, ordenados de menor a mayor.

3310 3355 3450 | 3480 3480 3490 | 3520 3540 3550 | 3650 3730 3925

$Q_1 = 3465$ $Q_2 = 3505$ (Mediana) $Q_3 = 3600$

La media, que es 3505 y los cuartiles $Q_1 = 3465$ y $Q_3 = 3600$ se calcularon ya en la sección 3.1. Si revisa los datos encontrará que el valor menor es 3310 y el valor mayor es 3925. Así, el resumen de cinco números correspondiente a los datos de los salarios iniciales es 3310, 3465, 3505, 3600, 3925. Entre cada dos números adyacentes del resumen de cinco números se encuentran aproximadamente 25% de los datos.

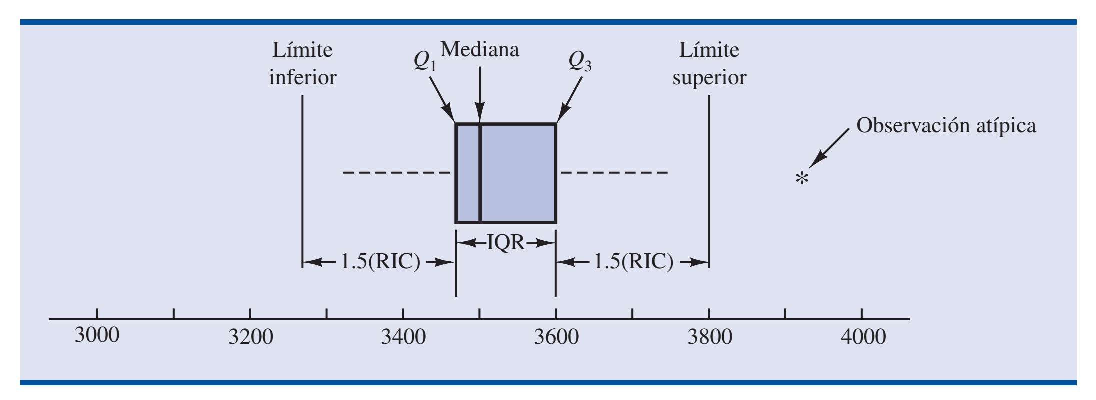
Diagrama de caja

Un **diagrama de caja** es un resumen gráfico de los datos con base en el resumen de cinco números. La clave para la elaboración de un diagrama de caja es el cálculo de la mediana y de los cuartiles Q_1 y Q_3 También se necesita el rango intercuartílico, RIC $= Q_3 - Q_1$. En la figura 3.5 se presenta el diagrama de caja de los datos de los salarios mensuales iniciales. Los pasos para elaborar un diagrama de caja son los siguientes.

Los diagramas de caja proporcionan otra manera de identificar observaciones atípicas. Pero no necesariamente se identifican los mismos valores que los correspondientes a un punto z menor que −3 o mayor que +3. Puede emplear cualquiera de estos procedimientos, o los dos.

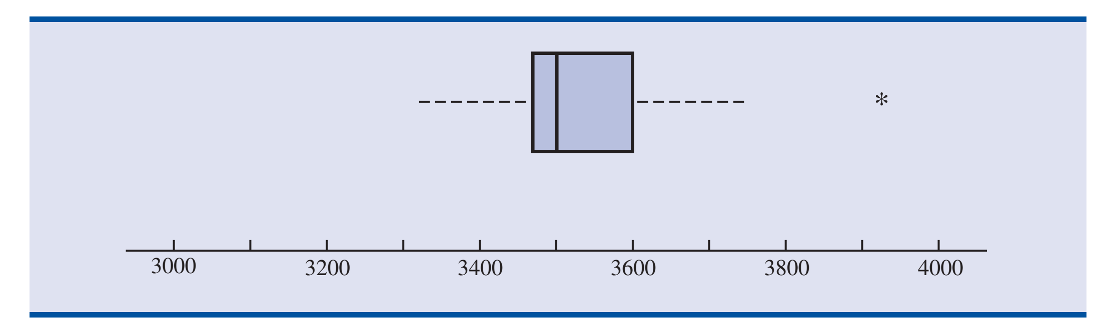
1. Se dibuja una caja cuyos extremos se localicen en el primer y tercer cuartiles. En los datos de los salarios iniciales $Q_1 = 3465$ y $Q_3 = 3600$. Esta caja contiene 50% de los datos centrales.
2. En el punto donde se localiza la mediana (3505 en los datos de los salarios) se traza una línea vertical.
3. Usando el rango intercuartílico, RIC $= Q_3 - Q_1$, se localizan los *límites*. En un diagrama de caja los límites se encuentran 1.5(RIC) abajo del Q_1 y 1.5(RIC) arriba del Q_3. En el caso de los salarios, RIC $= Q_3 - Q_1 = 3600 - 3465 = 135$. Por tanto, los límites son $3465 - 1.5(135) = 3262.5$ y $3600 + 1.5(135) = 3802.5$. Los datos que quedan fuera de estos límites se consideran *observaciones atípicas*.
4. A las líneas punteadas que se observan en la figura 3.5 se les llama *bigotes*. Los bigotes van desde los extremos de la caja hasta los valores menor y mayor *de los límites* calculados en el paso 3. Por tanto, los bigotes terminan en los salarios cuyos valores son 3310 y 3730.
5. Por último mediante un asterisco se indica la localización de las observaciones atípicas. En la figura 3.5 se observa que hay una observación atípica, 3925.

FIGURA 3.5 DIAGRAMA DE CAJA DE LOS SALARIOS INICIALES, EN EL QUE SE MUESTRAN LAS LÍNEAS QUE INDICAN LOS LÍMITES INFERIOR Y SUPERIOR



En la figura 3.5 se incluyeron las líneas que indican la localización de los límites superior e inferior. Estas líneas se dibujaron para mostrar cómo se calculan los límites y dónde se localizan en los datos de los salarios iniciales. Los límites, aunque siempre se calculan, por lo general no se dibujan en el diagrama de caja. En la figura 3.6 se muestra la apariencia usual del diagrama de caja de los datos de los salarios iniciales.

FIGURA 3.6 DIAGRAMA DE CAJA DE LOS DATOS DE LOS SALARIOS INICIALES



NOTAS Y COMENTARIOS

1. Una ventaja de los procedimientos del análisis exploratorio de datos es que son fáciles de usar; son necesarios pocos cálculos. Simplemente se ordenan los datos de menor a mayor y se identifican los cinco números del resumen de cinco números. Después se construye el diagrama de caja. No es necesario calcular la media ni la desviación estándar de los datos.
2. En el apéndice 3.1 se muestra cómo elaborar el diagrama de caja de los datos de los salarios iniciales empleando Minitab. El diagrama de caja que se obtiene es similar al de la figura 3.6, pero puesto de lado.

Ejercicios

Métodos

36. Considere una muestra cuyos valores son 27, 25, 20, 15, 30, 34, 28 y 25. Dé el resumen de cinco números de estos datos
37. Muestre diagrama de caja para los datos del ejercicio 36.
38. Elabore el resumen de cinco números y el diagrama de caja de los datos: 5, 15, 18, 10, 8, 12, 16, 10, 6.



39. En un conjunto de datos, el primer cuartil es 42 y el tercer cuartil es 50. Calcule los límites inferior y superior del diagrama de caja correspondiente. El dato con el valor 65, ¿debe considerarse como una observación atípica?

Aplicaciones

40. Ebby Halliday Realtors suministra publicidad sobre propiedades exclusivas ubicadas en Estados Unidos. A continuación se dan los precios de 22 propiedades (*The Wall Street Journal*, 16 de enero de 2004). Los precios se dan en miles



Property

1500	700	2995
895	619	880
719	725	3100
619	739	1699
625	799	1120
4450	2495	1250
2200	1395	912
1280		

a. Muestre el resumen de cinco números.
b. Calcule los límites inferior y superior.
c. La propiedad de mayor precio, $4 450 000, domina el lago White Rock en Dallas, Texas. ¿Esta propiedad se puede considerar como un valor atípico? Explique.
d. La segunda propiedad más cara que aparece en la lista es de $3 100 000, ¿debe considerarse como valor atípico? Explique.
e. Dibuje el diagrama de caja.



41. A continuación se presentan las ventas, en millones de dólares, de 21 empresas farmacéuticas.

8 408	1 374	1872	8879	2459	11 413
608	14 138	6452	1850	2818	1 356
10 498	7 478	4019	4341	739	2 127
3 653	5 794	8305			

a. Proporcione el resumen de cinco números.
b. Calcule los límites superior e inferior.
c. ¿Hay alguna observación atípica en estos datos?
d. Las ventas de Johnson & Johnson son las mayores de la lista, $14 138 millones. Suponga que se comete un error al registrar los datos (un error de transposición) y en lugar del valor dado se registra $41 138 millones. ¿Podría detectar este problema con el método de detección de observaciones atípicas del inciso c, de manera que se pudiera corregir este dato?
e. Dibuje el diagrama de caja.

42. Las nóminas en la liga mayor de béisbol siguen aumentando. Las nóminas de los equipos, en millones, son las siguientes (*USA Today* Online Database, marzo de 2006).



Equipo	Nómina	Equipo	Nómina
Arizona	$ 62	Milwaukee	$ 40
Atlanta	86	Minnesota	56
Baltimore	74	NY Mets	101
Boston	124	NY Yankees	208
Chi Cubs	87	Oakland	55
Chi White Sox	75	Philadelphia	96
Cincinnati	62	Pittsburgh	38
Cleveland	42	San Diego	63
Colorado	48	San Francisco	90
Detroit	69	Seattle	88
Florida	60	St. Louis	92
Houston	77	Tampa Bay	30
Kansas City	37	Texas	56
LA Angels	98	Toronto	46
LA Dodgers	83	Washington	49

a. ¿Cuál es la mediana de la nómina?
b. Proporcione el resumen de cinco números.
c. ¿Es una observación atípica la nómina de $208 millones de los Yankees de Nueva York? Explique.
d. Dibuje un diagrama de caja.

43. El presidente de la Bolsa de Nueva York, Richard Grasso, y su junta directiva se vieron cuestionados por el gran paquete de compensaciones pagado a Grasso. El salario más bonos de Grasso, $8.5 millones, superó el de todos los altos ejecutivos de las principales empresas de servicios financieros. Los datos siguientes muestran los salarios anuales más bonos pagados a los altos eje-

cutivos de 14 empresas de servicios financieros (*The Wall Street Journal*, 17 de septiembre de 2003). Los datos se dan en millones.

Empresa	Salario/bono	Empresa	Salario/bono
Aetna	$3.5	Fannie Mae	$4.3
AIG	6.0	Federal Home Loan	0.8
Allstate	4.1	Fleet Boston	1.0
American Express	3.8	Freddie Mac	1.2
Chubb	2.1	Mellon Financial	2.0
Cigna	1.0	Merrill Lynch	7.7
Citigroup	1.0	Wells Fargo	8.0

a. ¿Cuál es la mediana del salario más bono pagado a los altos ejecutivos de las 14 empresas de servicios financieros?
b. Obtenga el resumen de cinco números.
c. ¿Se debe considerar el salario más bonos de Grasso, $8.5 millones, como una observación atípica en el grupo de altos ejecutivos? Explique.
d. Presente el diagrama de caja.



44. En la tabla 3.6 se presentan 46 fondos mutualistas y sus rendimientos porcentuales anuales. (*Smart Money,* febrero de 2004.)
    a. ¿Cuáles son los rendimientos porcentuales promedio y la mediana de estos fondos mutualistas?
    b. ¿Cuáles son el primer y tercer cuartil?
    c. Obtenga el resumen de cinco números.
    d. ¿Hay alguna observación atípica en estos datos? Presente el diagrama de caja.

TABLA 3.6 RENDIMIENTOS PORCENTUALES ANUALES EN FONDOS MUTUALISTAS



Fondo mutualista	Rendimiento (%)	Fondo mutualista	Rendimiento (%)
Alger Capital Appreciation	23.5	Nations Small Company	21.4
Alger LargeCap Growth	22.8	Nations SmallCap Index	24.5
Alger MidCap Growth	38.3	Nations Strategic Growth	10.4
Alger SmallCap	41.3	Nations Value Inv	10.8
AllianceBernstein Technology	40.6	One Group Diversified Equity	10.0
Federated American Leaders	15.6	One Group Diversified Int'l	10.9
Federated Capital Appreciation	12.4	One Group Diversified Mid Cap	15.1
Federated Equity-Income	11.5	One Group Equity Income	6.6
Federated Kaufmann	33.3	One Group Int'l Equity Index	13.2
Federated Max-Cap Index	16.0	One Group Large Cap Growth	13.6
Federated Stock	16.9	One Group Large Cap Value	12.8
Janus Adviser Int'l Growth	10.3	One Group Mid Cap Growth	18.7
Janus Adviser Worldwide	3.4	One Group Mid Cap Value	11.4
Janus Enterprise	24.2	One Group Small Cap Growth	23.6
Janus High-Yield	12.1	PBHG Growth	27.3
Janus Mercury	20.6	Putnam Europe Equity	20.4
Janus Overseas	11.9	Putnam Int'l Capital Opportunity	36.6
Janus Worldwide	4.1	Putnam International Equity	21.5
Nations Convertible Securities	13.6	Putnam Int'l New Opportunity	26.3
Nations Int'l Equity	10.7	Strong Advisor Mid Cap Growth	23.7
Nations LargeCap Enhd. Core	13.2	Strong Growth 20	11.7
Nations LargeCap Index	13.5	Strong Growth Inv	23.2
Nation MidCap Index	19.5	Strong Large Cap Growth	14.5



3.5 Medidas de la asociación entre dos variables

Hasta ahora se han examinado métodos numéricos que resumen datos en *una sola variable*. Con frecuencia los administradores o quienes toman decisiones necesitan conocer la *relación entre dos variables.* En esta sección se presentan la covarianza y la correlación como medidas descriptivas de la relación entre dos variables.

Se empieza retomando la aplicación concerniente a la tienda de equipos de sonido que se presentó en la sección 2.4. El administrador de la tienda desea determinar la relación entre el número de comerciales televisados en un fin de semana y las ventas de la tienda durante la semana siguiente. En la tabla 3.7 se presentan datos muestrales de las ventas expresadas en cientos de dólares. En esta tabla se presentan 10 observaciones ($n = 10$), una por cada semana. El diagrama de dispersión en la figura 3.7 muestra una relación positiva, en que las mayores ventas (y) están asociadas con mayor número de comerciales (x). En efecto, el diagrama de dispersión sugiere que podría emplearse una línea recta como aproximación a esta relación. En la argumentación siguiente se introduce la **covarianza** como una medida descriptiva de la asociación entre dos variables.

Covarianza

En una muestra de tamaño n con observaciones (x_1, y_1), (x_2, y_2), etc., la covarianza muestral se define como sigue:

COVARIANZA MUESTRAL

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad \textbf{(3.10)}$$

Esta fórmula aparea cada x_i con una y_i. Después se suman los productos obtenidos al multiplicar la desviación de cada x_i de su media muestral $\bar{x}$ por la desviación de la y_i correspondiente de su media muestral $\bar{y}$; esta suma se divide entre $n - 1$.

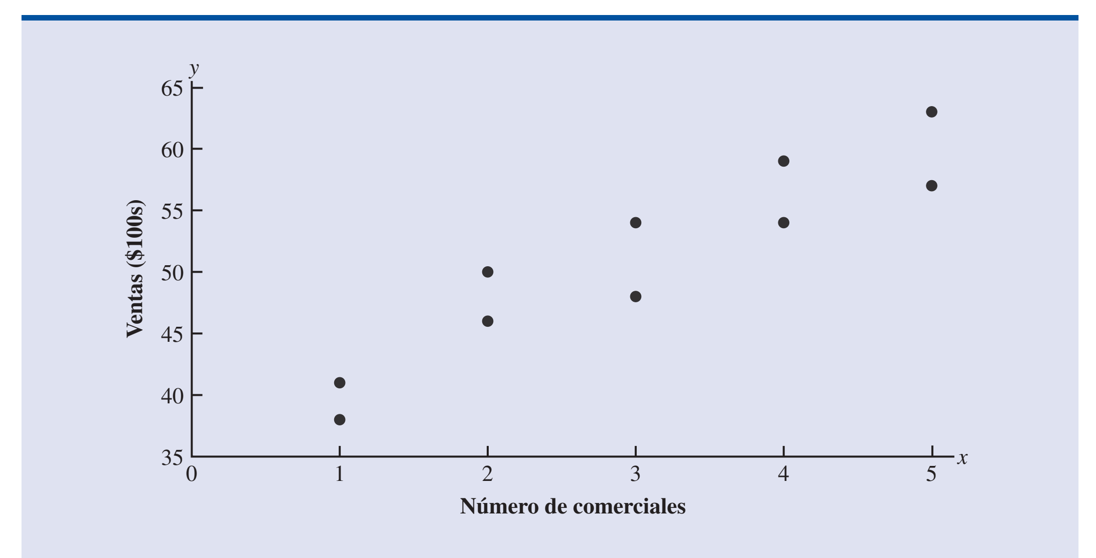
TABLA 3.7 DATOS MUESTRALES DE LA TIENDA DE EQUIPOS DE SONIDO

Semana	Número de comerciales x	Volumen de ventas ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46



Stereo

FIGURA 3.7 DATOS MUESTRALES DE LA TIENDA DE EQUIPOS DE SONIDO



Para medir, en el problema de la tienda de equipo de sonido, la fuerza de la relación lineal entre el número de comerciales x y el volumen de ventas y, se usa la ecuación (3.10) para calcular la covarianza muestral. En la tabla 3.8 se muestra el cálculo de $\Sigma(x_i - \bar{x})(y_i - \bar{y})$. Observe que $\bar{x} = 30/10 = 3$ y $\bar{y} = 510/10 = 51$. Empleando la ecuación (3.10) se encuentra que la covarianza muestral es

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{9} = 11$$

TABLA 3.8 CÁLCULO DE LA COVARIANZA MUESTRAL

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totales	30	510	0	0	99

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{10 - 1} = 11$$

La fórmula para calcular la covarianza de una población de tamaño N es semejante a la ecuación (3.10), pero la notación usada es diferente para indicar que se está trabajando con toda la población.

COVARIANZA POBLACIONAL

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N} \quad \textbf{(3.11)}$$

En la ecuación (3.11) μ_x se usa para denotar la media poblacional de la variable x y μ_y para denotar la media poblacional de la variable y. La covarianza σ_{xy} está definida para una población de tamaño N.

Interpretación de la covarianza

Para ayudar a la interpretación de la covarianza muestral, considere la figura 3.8; presenta el mismo diagrama de dispersión de la figura 3.7 pero con una línea vertical punteada en $\bar{x} = 3$ y una línea horizontal punteada en $\bar{y} = 51$. Estas líneas dividen a la gráfica en cuatro cuadrantes. Los puntos del cuadrante I corresponden a x_i mayor que $\bar{x}$ y y_i mayor que $\bar{y}$, los puntos del cuadrante II corresponden a x_i menor que $\bar{x}$ y y_i mayor que $\bar{y}$, etc. Por tanto, los valores de $(x_i - \bar{x})(y_i - \bar{y})$ serán positivos para los puntos del cuadrante I, negativos para los puntos del cuadrante II, positivos para los puntos del cuadrante III y negativos para los puntos del cuadrante IV.

La covarianza es una medida de la asociación lineal entre dos variables.

Si el valor de s_{xy} es positivo, los puntos que más influyen sobre s_{xy} deberán encontrarse en los cuadrantes I y III. Por tanto, s_{xy} positivo indica que hay una asociación lineal positiva entre x y y; es decir, que a medida que el valor de x aumenta, el valor de y aumenta. Si s_{xy} es negativo, los puntos que más influyen sobre s_{xy} deberán encontrarse en los cuadrantes II y IV. Entonces, s_{xy} negativo indica que hay una asociación lineal negativa entre x y y; esto es, conforme el valor de x aumenta, el valor de y disminuye. Por último, si los puntos tienen distribución uniforme en los cuatro cuadrantes, s_{xy} tendrá un valor cercano a cero, lo que indicará que no hay asociación lineal entre x y y. En la figura 3.9 se muestran los valores de s_{xy} esperables en tres tipos de diagramas de dispersión.

FIGURA 3.8 DIAGRAMA DE DISPERSIÓN DIVIDIDO PARA LA TIENDA DE EQUIPOS DE SONIDO

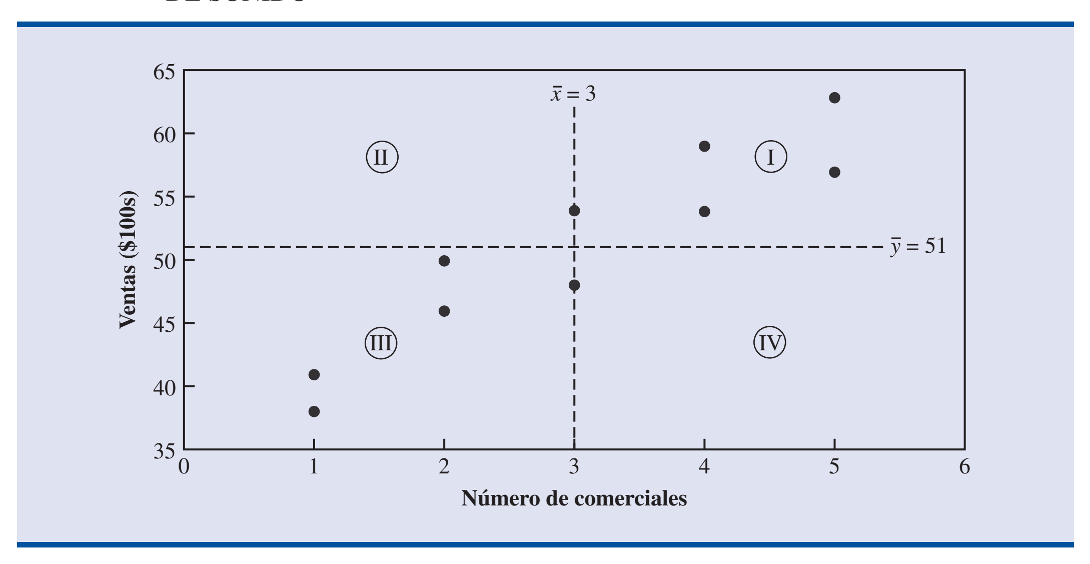
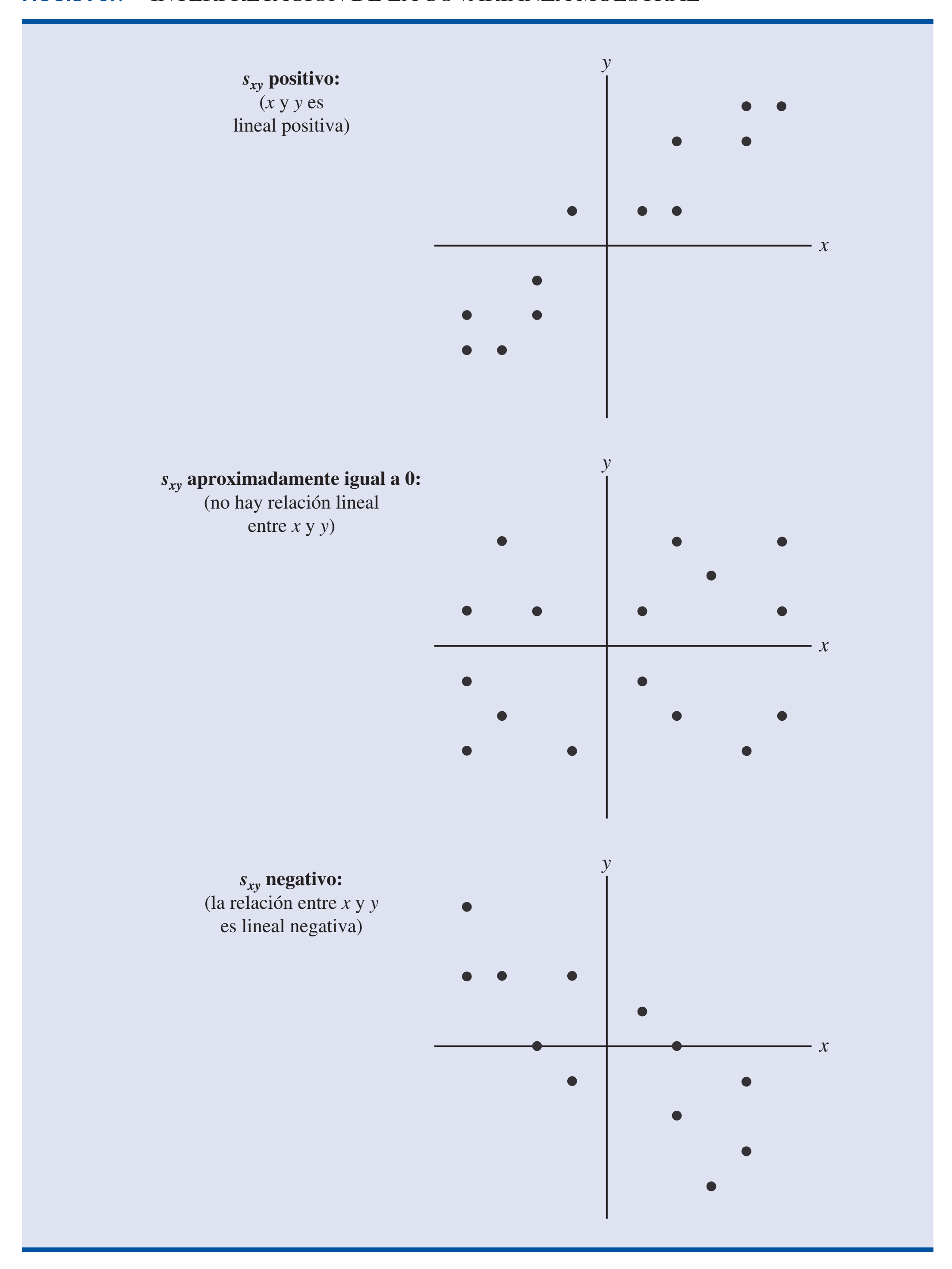


FIGURA 3.9 INTERPRETACIÓN DE LA COVARIANZA MUESTRAL



Si observa otra vez la figura 3.8, encontrará que el diagrama de dispersión de la tienda de equipos de sonido tiene un patrón similar a la gráfica superior de la figura 3.9. Como es de esperarse, el valor de la covarianza muestral indica que hay una relación lineal positiva en la que $s_{xy} = 11$.

Por la argumentación anterior parece que un valor positivo grande de la varianza indica una relación lineal positiva fuerte y que un valor negativo grande indica una relación lineal negativa fuerte. Sin embargo, un problema en el uso de la covarianza, como medida de la fuerza de la relación lineal, es que el valor de la covarianza depende de las unidades de medición empleadas para x y y. Suponga, por ejemplo, que se desea medir la relación entre la estatura x y el peso y de las personas. Es claro que la fuerza de la relación deberá ser la misma, ya sea que la altura se mida en pies o en pulgadas. Sin embargo, cuando la estatura se mide en pulgadas, los valores de $(x_i - \bar{x})$ son mayores que cuando se mide en pies. En efecto, cuando la estatura se mide en pulgadas, el valor del numerador $\Sigma(x_i - \bar{x})(y_i - \bar{y})$ de la ecuación (3.10) es mayor —entonces la covarianza es mayor— siendo que en realidad la relación no varía. Una medida de la relación entre dos variables, a la cual no le afectan las unidades de medición empleadas para x y y, es el **coeficiente de correlación**.

Coeficiente de correlación

Para datos muestrales el coeficiente de correlación del producto–momento de Pearson está definido como sigue.

COEFICIENTE DE CORRELACIÓN DEL PRODUCTO–MOMENTO DE PEARSON: DATOS MUESTRALES

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad \textbf{(3.12)}$$

donde

r_{xy} = coeficiente de correlación muestral
s_{xy} = covarianza muestral
s_x = desviación estándar muestral de x
s_y = desviación estándar muestral de y

En la ecuación (3.12) se observa que el coeficiente de correlación del producto–momento de Pearson para datos muestrales (llamado *coeficiente de correlación muestral*) se calcula dividiendo la covarianza muestral entre el producto de la desviación estándar muestral de x por la desviación estándar muestral de y.

A continuación se calcula el coeficiente de correlación de los datos de la tienda de equipos para sonido. A partir de la tabla 3.8, se calcula la desviación estándar muestral de las dos variables.

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{20}{9}} = 1.49$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{566}{9}} = 7.93$$

Ahora, como $s_{xy} = 11$, el coeficiente de correlación muestral es igual a

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{(1.49)(7.93)} = +0.93$$

La fórmula para calcular el coeficiente de correlación de una población que se denota con la letra griega ρ_{xy} (ro) es la siguiente.

COEFICIENTE DE CORRELACIÓN DEL PRODUCTO–MOMENTO DE PEARSON: DATOS POBLACIONALES

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad \textbf{(3.13)}$$

donde

ρ_{xy} = coeficiente de correlación poblacional
σ_{xy} = covarianza poblacional
σ_x = desviación estándar poblacional de x
σ_y = desviación estándar poblacional de y

El coeficiente de correlación muestral r_{xy} proporciona un estimador del coeficiente de correlación poblacional ρ_{xy}.

El coeficiente de correlación muestral r_{xy} proporciona un estimador del coeficiente de correlación poblacional ρ_{xy}.

Interpretación del coeficiente de correlación

Primero se considerará un ejemplo sencillo que ilustra el concepto de una relación lineal positiva perfecta. En el diagrama de dispersión en la figura 3.10 se representa la relación entre x y y con base en los datos muestrales siguientes.

x_i	y_i
5	10
10	30
15	50

FIGURA 3.10 DIAGRAMA DE DISPERSIÓN QUE REPRESENTA UNA RELACIÓN LINEAL POSITIVA PERFECTA

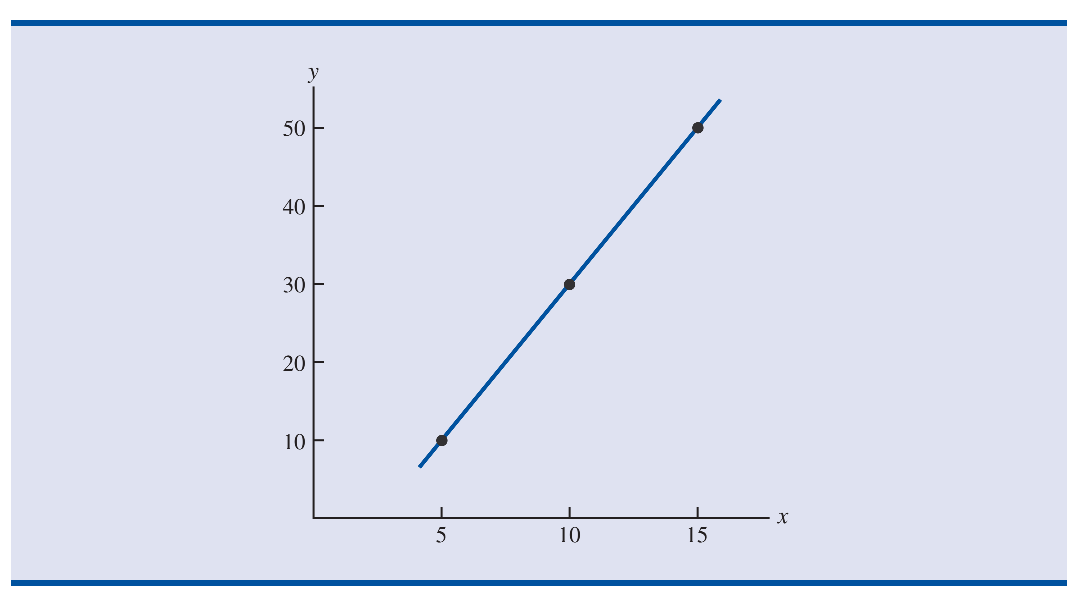


TABLA 3.9 CÁLCULOS PARA OBTENER EL COEFICIENTE DE CORRELACIÓN MUESTRAL

	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	5	10	−5	25	−20	400	100
	10	30	0	0	0	0	0
	15	50	5	25	20	400	100
Totales	30	90	0	50	0	800	200
	$\bar{x} = 10$	$\bar{y} = 30$					

La línea recta trazada a través de los tres puntos expresa una relación lineal perfecta entre x y y. Para emplear la ecuación (3.12) en el cálculo de la correlación muestral, es necesario calcular primero s_{xy}, s_x y s_y. En la tabla 3.9 se muestran parte de los cálculos. Con los resultados de la tabla 3.9 se tiene

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{200}{2} = 100$$

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{50}{2}} = 5$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{800}{2}} = 20$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{100}{5(20)} = 1$$

El coeficiente de correlación va desde −1 hasta +1. Los valores cercanos a −1 o a +1 corresponden a una relación lineal fuerte. Entre más cercano a cero sea el valor de la correlación, más débil es la relación lineal.

De manera que el valor del coeficiente de correlación muestral es 1.

En general, puede demostrar que si todos los valores del conjunto de datos caen en una línea recta con pendiente positiva, el coeficiente de correlación será +1; es decir, un coeficiente de correlación de +1 corresponde a una relación lineal positiva perfecta entre x y y. Por otra parte, si los puntos del conjunto de datos caen sobre una línea recta con pendiente negativa, el coeficiente de correlación muestral será −1; un coeficiente de correlación de −1 corresponde a una relación lineal negativa perfecta entre x y y.

Suponga ahora que un conjunto de datos muestra una relación lineal positiva entre x y y, pero que la relación no es perfecta. El valor de r_{xy} será menor a 1, indicando que no todos los puntos del diagrama de dispersión se encuentran en una línea recta. Entre más se desvíen los puntos de una relación lineal positiva perfecta, más pequeño será r_{xy}. Si r_{xy} es igual a cero, entonces no hay relación lineal entre x y y; si r_{xy} tiene un valor cercano a cero, la relación lineal es débil.

Recuerde que en el caso de los datos de la tienda de equipo de sonido $r_{xy} = +0.93$. Entonces se concluye que existe una relación lineal fuerte entre el número de comerciales y las ventas. Más en específico, un aumento en el número de comerciales se asocia con un incremento en las ventas.

Para terminar, es preciso destacar que la correlación proporciona una medida de la asociación lineal y no necesariamente de la causalidad. Que la correlación entre dos variables sea alta no significa que los cambios en una de las variables ocasionen modificaciones en la otra. Por ejemplo, quizá encuentre que las evaluaciones de la calidad y los precios de los restaurantes tengan una correlación positiva. Sin embargo, aumentar los precios de un restaurante no hará que las evaluaciones mejoren.

Ejercicios

Métodos



45. Las siguientes son cinco observaciones de dos variables

x_i	4	6	11	3	16
y_i	50	50	40	60	30

a. Elabore un diagrama de dispersión con x en el eje horizontal.
b. ¿Qué indica el diagrama de dispersión elaborado en el inciso a respecto a la relación entre las dos variables?
c. Calcule e interprete la covarianza muestral.
d. Calcule e interprete el coeficiente de correlación muestral.

46. Las siguientes son cinco observaciones de dos variables.

x_i	6	11	15	21	27
y_i	6	9	6	17	12

a. Elabore un diagrama de dispersión con estas variables.
b. ¿Qué indica este diagrama de dispersión respecto de la relación entre x y y?
c. Calcule e interprete la covarianza muestral.
d. Calcule e interprete el coeficiente de correlación muestral.

Aplicaciones

47. Nielsen Media Research proporciona dos medidas de la audiencia que tienen los programas de televisión: un *rating* de los programas, porcentaje de hogares que tienen televisión y están viendo determinado programa, y un *share* de los programas de televisión, porcentaje de hogares que tienen la televisión encendida y están viendo un determinado programa. Los datos siguientes muestran los datos de *rating* y *share* de Nielsen para la final de la liga mayor de básquetbol en un periodo de nueve años. (Associated Press, 27 de octubre de 2003).

Rating	19	17	17	14	16	12	15	12	13
Share	32	28	29	24	26	20	24	20	22

a. Elabore un diagrama de dispersión con los *ratings* en el eje horizontal.
b. ¿Cuál es la relación entre *rating* y *share*? Explique.
c. Calcule e interprete la covarianza muestral.
d. Calcule el coeficiente de correlación muestral. ¿Qué dice este valor acerca de la relación entre *rating* y *share*?

48. En un estudio del departamento de transporte sobre la velocidad y el rendimiento de la gasolina en automóviles de tamaño mediano se obtuvieron los datos siguientes.

Velocidad	30	50	40	55	30	25	60	25	50	55
Rendimiento	28	25	25	23	30	32	21	35	26	25

Calcule e interprete el coeficiente de correlación muestral.

49. *PC World* proporciona evaluaciones de 15 *notebook* PCs (*PC World*, febrero de 2000). La puntuación de funcionamiento mide cuán rápido corre una PC un conjunto de aplicaciones usadas en administración, en comparación con una máquina de línea base. Por ejemplo una PC cuya puntuación de funcionamiento es 200 es dos veces más rápida que una máquina de línea base. Para proporcionar una evaluación general de cada *notebook* probada en el estudio se empleó una escala de 100 puntos. Una puntuación general alrededor de 90 es excepcional, mientras que una de 70 es buena. En la tabla 3.10 se muestran las puntuaciones de funcionamiento y las puntuaciones generales de 15 *notebooks*.

TABLA 3.10 PUNTUACIONES DE FUNCIONAMIENTO Y PUNTUACIONES GENERALES DE 15 *NOTEBOOK* PC



Notebook	Puntuación de funcionamiento	Puntuación general
AMS Tech Roadster 15CTA380	115	67
Compaq Armada M700	191	78
Compaq Prosignia Notebook 150	153	79
Dell Inspiron 3700 C466GT	194	80
Dell Inspiron 7500 R500VT	236	84
Dell Latitude Cpi A366XT	184	76
Enpower ENP-313 Pro	184	77
Gateway Solo 9300LS	216	92
HP Pavilion Notebook PC	185	83
IBM ThinkPad I Series 1480	183	78
Micro Express NP7400	189	77
Micron TransPort NX PII-400	202	78
NEC Versa SX	192	78
Sceptre Soundx 5200	141	73
Sony VAIO PCG-F340	187	77

a. Calcule el coeficiente de correlación muestral.

b. ¿Qué indica el coeficiente de correlación muestral acerca de la relación entre la puntuación de funcionamiento y la puntuación general?

50. El Promedio Industrial Dow Jones (DJIA, por sus siglas en inglés) y el Standard & Poor's 500 Index (S&P 500) se usan para medir el mercado bursátil. El DJIA se basa en el precio de las acciones de 30 empresas grandes; el S&P 500 se basa en los precios de las acciones de 500 empresas. Si ambas miden el mercado bursátil, ¿cuál es la relación entre ellas? En los datos siguientes se muestra el aumento porcentual diario o la disminución porcentual diaria del DJIA y del S&P 500 en una muestra de nueve días durante tres meses (*The Wall Street Journal*, 15 de enero a 10 de marzo de 2006).



DJIA	0.20	0.82	−0.99	0.04	−0.24	1.01	0.30	0.55	−0.25
S&P 500	0.24	0.19	−0.91	0.08	−0.33	0.87	0.36	0.83	−0.16

a. Muestre el diagrama de dispersión.

b. Calcule el coeficiente de correlación muestral de estos datos.

c. Discuta la asociación entre DJIA y S&P 500. ¿Es necesario consultar ambos para tener una idea general sobre el mercado bursátil diario?

51. Las temperaturas más altas y más bajas en 12 ciudades de Estados Unidos son las siguientes. (Weather Channel, 25 de enero de 2004.)



Ciudad	Alta	Baja	Ciudad	Alta	Baja
Albany	9	−8	Los Angeles	62	47
Boise	32	26	New Orleans	71	55
Cleveland	21	19	Portland	43	36
Denver	37	10	Providence	18	8
Des Moines	24	16	Raleigh	28	24
Detroit	20	17	Tulsa	55	38

a. ¿Cuál es la media muestral de las temperaturas diarias más elevadas?
b. ¿Cuál es la media muestral de las temperaturas diarias más bajas?
c. ¿Cuál es la correlación entre temperaturas más elevadas y temperaturas más bajas?



La media ponderada y el empleo de datos agrupados

En la sección 3.1 se presentó la media como una de las medidas más importantes de localización central. La fórmula para la media de una muestra en la que hay n observaciones se escribe como sigue.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad \textbf{(3.14)}$$

En esta fórmula, a cada x_i se le da la misma importancia o el mismo peso. Aunque esto es lo más común, en algunas situaciones la media se calcula dando a cada observación un peso que refleja su importancia. A una media calculada de esta manera se le llama **media ponderada**.

Media ponderada

La media ponderada se calcula:

MEDIA PONDERADA

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} \quad \textbf{(3.15)}$$

donde

x_i = valor de la observación i
w_i = peso para la observación i

Si los datos provienen de una muestra, la ecuación (3.15) proporciona la media ponderada muestral. Si son de una población, μ se sustituye por $\bar{x}$ en la ecuación (3.15) y se obtiene la media ponderada poblacional.

Como ejemplo de la necesidad de la media ponderada muestral, considere la muestra siguiente de cinco compras de materia prima realizadas en los últimos tres meses.

Compra	Costo por libra ($)	Número de libras
1	3.00	1200
2	3.40	500
3	2.80	2750
4	2.90	1000
5	3.25	800

Observe que el costo por libra varía desde $2.80 hasta $3.40 y la cantidad comprada varía desde 500 hasta 2 750 libras. Suponga que el administrador quiere información sobre el costo medio por libra de la materia prima. Como las cantidades compradas varían, es necesario emplear la fórmula para la media ponderada. Los valores de los datos de los cinco costos por libra son $x_1 = 3.00$, $x_2 = 3.40$, $x_3 = 2.80$, $x_4 = 2.90$, y $x_5 = 3.25$. El costo medio ponderado por libra se ob-

tiene ponderando cada costo con su cantidad correspondiente. Por ejemplo, los pesos (de ponderación) son $w_1 = 1200$, $w_2 = 500$, $w_3 = 2750$, $w_4 = 1000$ y $w_5 = 800$. De acuerdo con la ecuación (3.15) la media ponderada se calcula:

$$\bar{x} = \frac{1200(3.00) + 500(3.40) + 2750(2.80) + 1000(2.90) + 800(3.25)}{1200 + 500 + 2750 + 1000 + 800}$$

$$= \frac{18\,500}{6250} = 2.96$$

Así, los cálculos de la media ponderada indican que el costo medio por libra de materia prima es \$2.96. Observe que si hubiera usado la ecuación (3.14) en lugar de la fórmula para la media ponderada, hubiera obtenido resultados engañosos. En ese caso la media de los valores de los cinco costos por libra sería (3.00 + 3.40 + 2.80 + 2.90 + 3.25)/5 = 15.35/5 = \$3.07, valor que exagera el costo medio real por libra comprada.

La selección de las ponderaciones para el cálculo de una determinada media ponderada dependen de la aplicación. Un ejemplo muy conocido por los estudiantes es el promedio de las calificaciones (en Estados Unidos). En este caso los valores de los datos son 4 que corresponde a A, 3 que corresponde a B, 2 que corresponde a C, 1 que corresponde a D y 0 que corresponde a F. Los pesos son los créditos por hora de cada materia. El ejercicio 54 al final de esta sección es un ejemplo del cálculo de esta media ponderada. En otros cálculos de la media ponderada se emplean como pesos cantidades como libras, dólares o volumen. En cualquier caso, si la importancia de las observaciones varía, el analista debe elegir los pesos que mejor reflejen la relevancia de cada observación en la determinación de la media.

El cálculo de las calificaciones es un buen ejemplo del uso de la media ponderada.

Datos agrupados

En la mayor parte de los casos, las medidas de localización y variabilidad se calculan mediante los valores individuales de los datos. Sin embargo, otras veces sólo se tienen datos agrupados o datos en una distribución de frecuencias. En la argumentación siguiente se muestra cómo usar la fórmula de la media ponderada para obtener aproximaciones a la media, la varianza y la desviación estándar de **datos agrupados**.

En la sección 2.2 se presentó una distribución de las duraciones en días en una muestra de auditorías de fin de año de una empresa pequeña de contadores públicos. La distribución de frecuencias de las duraciones de las auditorías que se obtuvo de una muestra de 20 clientes se presenta de nuevo en la tabla 3.11. Con base en esta distribución de frecuencias, ¿cuál es la media muestral de la duración de las auditorías?

Para calcular la media usando datos agrupados, considere el punto medio de cada clase como representativo de los elementos de esa clase. Si M_i denota el punto medio de la clase i y f_i denota la frecuencia de la clase i. Entonces la fórmula para la media ponderada (3.15) se usa con los valores de los datos denotados por M_i y los pesos dados por las frecuencias f_i. En este caso, el denominador de la ecuación (3.15) es la suma de las frecuencias, que es el tamaño de la muestra n.

TABLA 3.11 DISTRIBUCIÓN DE FRECUENCIAS DE LAS DURACIONES DE LAS AUDITORÍAS

Duración de la auditoría (en días)	Frecuencia
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

Es decir, $\Sigma f_i = n$. De manera que la ecuación para la media muestral de datos agrupados es la siguiente:

MEDIA MUESTRAL DE DATOS AGRUPADOS

$$\bar{x} = \frac{\Sigma f_i M_i}{n} \quad \textbf{(3.16)}$$

donde

M_i = punto medio de la clase i
f_i = frecuencia de la clase i
n = tamaño de la muestra

Como el punto medio de clase, M_i, se encuentra a la mitad entre los límites de clase, en tabla 3.11 el punto medio de la primera clase, 10–14, es $(10 + 14)/2 = 12$. En la tabla 3.12 se presentan los cinco puntos medios de clase y los cálculos de la media ponderada de los datos de la duración de las auditorías. Como puede ver, la media muestral de la duración de las auditorías es 19 días.

Para calcular la varianza de datos agrupados se emplea una versión ligeramente modificada de la fórmula para la varianza dada en la ecuación (3.5). En la ecuación (3.5) los cuadrados de las desviaciones de los datos respecto a la media muestral se escribieron como $(x_i - \bar{x})^2$. Pero cuando se tienen datos agrupados no se conocen los valores. En este caso, se considera el punto medio de clase, M_i, como representativo de los valores x_i de la clase correspondiente. Por tanto, los cuadrados de las desviaciones respecto a la media $(x_i - \bar{x})^2$ son sustituidos por $(M_i - \bar{x})^2$. Entonces, igual que en el cálculo de la media muestral de datos agrupados, pondere cada valor por la frecuencia de la clase, f_i. La suma de los cuadrados de las desviaciones respecto a la media de todos los datos se aproxima mediante $\Sigma f_i(M_i - \bar{x})^2$. En el denominador aparece el término $n - 1$ en lugar de n, con objeto de hacer que la varianza muestral sea un estimador de la varianza poblacional. Por consiguiente, la fórmula usada para obtener la varianza muestral de datos agrupados es:

VARIANZA MUESTRAL PARA DATOS AGRUPADOS

$$s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} \quad \textbf{(3.17)}$$

TABLA 3.12 CÁLCULO DE LA VARIANZA MUESTRAL CON LOS DATOS AGRUPADOS DE LAS DURACIONES DE LAS AUDITORÍAS

Duración de la auditoría (días)	Punto medio de clase (M_i)	Frecuencia (f_i)	f_iM_i
10–14	12	4	48
15–19	17	8	136
20–24	22	5	110
25–29	27	2	54
30–34	32	1	32
		20	380

$$\text{Media muestral } \bar{x} = \frac{\Sigma f_i M_i}{n} = \frac{380}{20} = 19 \text{ días}$$

TABLA 3.13 CÁLCULO DE LA VARIANZA MUESTRAL CON LOS DATOS AGRUPADOS DE LAS DURACIONES DE LAS AUDITORÍAS

Duración de la auditoría (días)	Punto medio de clase (M_i)	Frecuencia (f_i)	Desviación ($M_i - \bar{x}$)	Cuadrado de la desviación $(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
10–14	12	4	−7	49	196
15–19	17	8	−2	4	32
20–24	22	5	3	9	45
25–29	27	2	8	64	128
30–34	32	1	13	169	169
		20			570
					$\Sigma f_i(M_i - \bar{x})^2$

$$\text{Varianza muestral } s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} = \frac{570}{19} = 30$$

En la tabla 3.13 se presenta el cálculo de la varianza muestral de las duraciones de las auditorías a partir de los datos agrupados de la tabla 3.11, ahí la varianza muestral es 30.

La desviación estándar de datos agrupados es simplemente la raíz cuadrada de la varianza de los datos agrupados. La desviación estándar muestral de los datos de las duraciones de las auditorías es $s = \sqrt{30} = 5.48$.

Antes de terminar esta sección sobre el cálculo de medidas de localización y de dispersión de datos agrupados, debe observar que las fórmulas (3.16) y (3.17) son para muestras. El cálculo de las medidas poblacionales es semejante. A continuación se presentan las fórmulas para la media y la varianza poblacional de datos agrupados.

MEDIA POBLACIONAL DE DATOS AGRUPADOS

$$\mu = \frac{\Sigma f_i M_i}{N} \tag{3.18}$$

VARIANZA POBLACIONAL DE DATOS AGRUPADOS

$$\sigma^2 = \frac{\Sigma f_i(M_i - \mu)^2}{N} \tag{3.19}$$

NOTAS Y COMENTARIOS

Al calcular los estadísticos descriptivos de datos agrupados, se usan los puntos medios de clase para aproximar los valores de los datos de cada clase. Por tanto, los estadísticos descriptivos de datos agrupados aproximan los estadísticos descriptivos que se obtendrían si se usaran los datos originales. En consecuencia, es recomendable calcular los estadísticos descriptivos con los datos originales y no con los datos agrupados, siempre que sea posible.

Ejercicios

Métodos

52. Considere los datos siguientes con sus pesos correspondientes

x_i	Peso (w_i)
3.2	6
2.0	3
2.5	2
5.0	8

a. Calcule la media ponderada.
b. Calcule la media muestral de los cuatro valores de los datos sin los pesos. Observe la diferencia que hay entre los resultados obtenidos con los dos métodos.



53. Considere los datos muestrales de la distribución de frecuencia siguiente.

Clase	Punto medio	Frecuencia
3–7	5	4
8–12	10	7
13–17	15	9
18–22	20	5

a. Calcule la media muestral.
b. Calcule la varianza muestral y la desviación estándar muestral.

Aplicaciones



54. El promedio de calificaciones de los estudiantes de ciertas escuelas universitarias es el cálculo de una media ponderada. A las calificaciones se les dan los valores siguientes: A (4), B (3), C (2), D (1) y F (0). Después de un semestre de 60 horas de créditos, un estudiante obtuvo las calificaciones siguientes: A en 9 horas de crédito, B en 15 horas, C en 33 horas y D en 3 horas.
a. Calcule el promedio de calificaciones de este estudiante.
b. En esta universidad los estudiantes deben tener un promedio de 2.5 para poder seguir sus estudios. ¿Dicho estudiante podrá seguir sus estudios?

55. *Bloomberg Personal Finance* (julio/agosto de 2001) incluye las empresas siguientes en el portafolio de las inversiones que recomienda. A continuación se presentan las cantidades en dólares que asignan a cada acción en un portafolio con valor de $25 000.

Empresa	Portafolio ($)	Tasa de crecimiento estimado (%)	Rendimiento de dividendos (%)
Citigroup	3000	15	1.21
General Electric	5500	14	1.48
Kimberly-Clark	4200	12	1.72
Oracle	3000	25	0.00
Pharmacia	3000	20	0.96
SBC Communications	3800	12	2.48
WorldCom	2500	35	0.00

a. Use como pesos las cantidades en dólares del portafolio, ¿cuál es la tasa de crecimiento medio ponderado del portafolio?
b. ¿Cuál es el rendimiento medio ponderado de los dividendos en este portafolio?

56. En una investigación realizada entre los suscriptores de la revista *Fortune* se hizo la pregunta siguiente: "De los últimos números ¿cuántos ha leído?" Suponga que en la distribución de frecuencia siguiente se resumen las 500 respuestas.

Números leídos		Frecuencia
0		15
1		10
2		40
3		85
4		350
	Total	500

a. ¿Cuál es la cantidad media de los últimos números que han leído los suscriptores?
b. ¿Cuál es la desviación estándar en la cantidad de los últimos números que han leído los suscriptores?

57. La distribución de frecuencias siguiente muestra los precios de las 30 acciones del Promedio Industrial Dow Jones (*The Wall Street Journal*, 16 de enero de 2006).

Precio por acción	Frecuencia
\$20–29	7
\$30–39	6
\$40–49	6
\$50–59	3
\$60–69	4
\$70–79	3
\$80–89	1

Calcule el precio medio por acción y la desviación estándar de los precios por acción en el Promedio Industrial Dow Jones.

Resumen

En este capítulo se presentaron varios estadísticos descriptivos que sirven para resumir la localización, variabilidad y forma de la distribución de un conjunto de datos. A diferencia de los procedimientos gráficos y tabulares presentados en el capítulo 2, las medidas presentadas resumen los datos con valores numéricos. Cuando dichos valores numéricos se obtienen de una muestra, son llamados estadísticos muestrales, cuando se obtienen de una población, son parámetros poblacionales. A continuación se presenta la notación que se acostumbra emplear para estadísticos muestrales y para parámetros poblacionales.

En inferencia estadística a los estadísticos muestrales se les conoce como estimadores puntuales de los parámetros poblacionales.

	Estadístico muestral	Parámetro poblacional
Media	$\bar{x}$	μ
Varianza	s^2	σ^2
Desviación estándar	s	σ
Covarianza	s_{xy}	σ_{xy}
Correlación	r_{xy}	ρ_{xy}

Como medidas de localización central se definió la media, la mediana y la moda. Después se usó el concepto de percentiles para describir otras localizaciones en el conjunto de datos. A continuación se presentaron el rango, el rango intercuartílico, la varianza, la desviación estándar y el coeficiente de variación como medidas de variabilidad o de dispersión. La primera medida presentada para la forma de la distribución de los datos fue el sesgo; aquí, valores negativos corresponden a distribuciones de datos sesgadas a la izquierda, y valores positivos corresponden a distribuciones de datos sesgadas a la derecha. Después se describió cómo usar la media y la desviación estándar junto con el teorema de Chebyshev y la regla empírica para obtener más información acerca de la distribución de los datos y para identificar observaciones atípicas.

En la sección 3.4 se mostró cómo elaborar un resumen de cinco números y un diagrama de caja para obtener simultáneamente información sobre la localización, variabilidad y forma de una distribución. En la sección 3.5 se presentaron la covarianza y el coeficiente de correlación como medidas de la asociación entre dos variables. En la última sección se vio cómo calcular la media ponderada y cómo calcular media, varianza y desviación estándar de datos agrupados.

Los estadísticos descriptivos, aquí estudiados, pueden calcularse mediante paquetes de software para estadística y hojas de cálculo. En el apéndice 3.1 se muestra cómo obtener la mayor parte de estos estadísticos descriptivos usando Minitab. En el apéndice 3.2 se muestra el uso de Excel para los mismos propósitos.

Glosario

Estadístico muestral Valor numérico usado como una medida que resume una muestra (por ejemplo, la media muestral $\bar{x}$, la varianza muestral, s^2 y la desviación estándar muestral, s).

Parámetro poblacional Valor numérico que resume una población (por ejemplo, la media poblacional μ, la varianza poblacional, σ^2 y la desviación estándar poblacional, σ).

Estimador puntual Un estadístico muestral como $\bar{x}$, s^2 y s cuando se usa para estimar el parámetro poblacional correspondiente.

Media Medida de localización central que se calcula sumando los valores de los datos y dividiendo entre el número de observaciones.

Mediana Medida de localización central proporcionada por el valor central de los datos cuando éstos se han ordenado de menor a mayor.

Moda Medida de localización central, definida como el valor que se presenta con mayor frecuencia.

Percentil Un valor tal que por lo menos p por ciento de las observaciones son menores o iguales que este valor y por lo menos $(100 - p)$ por ciento de las observaciones son mayores o iguales que este valor. El percentil 50 es la mediana.

Cuartiles Los percentiles 25, 50 y 75, llamados cada uno primer cuartil, segundo cuartil (mediana) y tercer cuartil. Los cuartiles sirven para dividir al conjunto de datos en cuatro partes; cada una contiene aproximadamente 25% de los datos.

Rango Una medida de la variabilidad, que se define como el valor mayor menos el menor.

Rango intercuartílico (RIC) Una medida de la variabilidad, que se define como la diferencia entre el tercer y primer cuartil.

Varianza Una medida de la variabilidad que se basa en los cuadrados de las desviaciones de los datos respecto a la media.

Desviación estándar Una medida de variabilidad obtenida de la raíz cuadrada de la varianza.

Coeficiente de variación Medida de variabilidad relativa que se obtiene al dividir la desviación estándar entre la media y multiplicando el resultado por 100.

Sesgo Medida de la forma de la distribución de los datos. Datos sesgados a la izquierda tienen un sesgo negativo; una distribución de datos simétrica tiene sesgo cero, y datos sesgados a la derecha tienen sesgo positivo.

Punto z Valor que se calcula dividiendo la desviación respecto a la media $(x_i - \bar{x})$ entre la desviación estándar s. A los puntos z también se les conoce como valores estandarizados y denotan el número de desviaciones estándar que x_i se aleja de la media.
Teorema de Chebyshev Un teorema útil para obtener la proporción de valores en los datos que se encuentran a no más de un número determinado de desviaciones estándar de la media.
Regla empírica Regla empleada para calcular el porcentaje de los valores en los datos que se encuentran a no más de una, dos o tres desviaciones estándar de la media, cuando los datos muestran una distribución en forma de campana.
Observación atípica Datos que tienen un valor inusualmente grande o pequeño.
Resumen de cinco números Técnica para el análisis exploratorio de datos, usa cinco números para resumir los datos: el valor menor, el primer cuartil, la mediana, el tercer cuartil, y el valor mayor.
Diagrama de caja Resumen gráfico de los datos que se basa en el resumen de cinco números.
Covarianza Medida de la relación lineal entre dos variables. Si la covarianza es positiva, indica una relación positiva, y si es negativa, una relación negativa.
Coeficiente de correlación Medida de la relación lineal entre dos variables, que puede tener valores desde -1 hasta $+1$. Los valores cercanos a $+1$ indican una fuerte relación lineal positiva; valores cercanos a -1 muestran una fuerte relación lineal negativa, y valores cercanos a cero una ausencia de relación lineal.
Media ponderada Media que se obtiene asignando a cada uno de los valores un peso que refleja su importancia.
Datos agrupados Datos que se dan en intervalos de clase, como cuando se resumen para una distribución de frecuencias. No se tienen los valores de los datos originales.

Fórmulas clave

Media muestral

$$\bar{x} = \frac{\Sigma x_i}{n} \quad \textbf{(3.1)}$$

Media poblacional

$$\mu = \frac{\Sigma x_i}{N} \quad \textbf{(3.2)}$$

Rango intercuartílico

$$\text{RIC} = Q_3 - Q_1 \quad \textbf{(3.3)}$$

Varianza poblacional

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \quad \textbf{(3.4)}$$

Varianza muestral

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \quad \textbf{(3.5)}$$

Desviación estándar

$$\text{Desviación estándar muestral} = s = \sqrt{s^2} \quad \textbf{(3.6)}$$

$$\text{Desviación estándar poblacional} = \sigma = \sqrt{\sigma^2} \quad \textbf{(3.7)}$$

Coeficiente de variación

$$\left(\frac{\text{Desviación estándar}}{\text{Media}} \times 100\right)\% \quad \textbf{(3.8)}$$

Punto z

$$z_i = \frac{x_i - \bar{x}}{s} \quad \textbf{(3.9)}$$

Covarianza muestral

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad \textbf{(3.10)}$$

Covarianza poblacional

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N} \quad \textbf{(3.11)}$$

Coeficiente de correlación del producto–momento de Pearson: datos muestrales

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad \textbf{(3.12)}$$

Coeficiente de correlación del producto–momento de Pearson: datos poblacionales

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \textbf{(3.13)}$$

Media ponderada

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} \quad \textbf{(3.15)}$$

Media muestral de datos agrupados

$$\bar{x} = \frac{\Sigma f_i M_i}{n} \quad \textbf{(3.16)}$$

Varianza muestral de datos agrupados

$$s^2 = \frac{\Sigma f_i (M_i - \bar{x})^2}{n - 1} \quad \textbf{(3.17)}$$

Media poblacional de datos agrupados

$$\mu = \frac{\Sigma f_i M_i}{N} \quad \textbf{(3.18)}$$

Varianza poblacional de datos agrupados

$$\sigma^2 = \frac{\Sigma f_i (M_i - \mu)^2}{N} \quad \textbf{(3.19)}$$

Ejercicios complementarios

58. De acuerdo con 2003 Annual Consumer Spending Survey, el cargo promedio mensual a una tarjeta de crédito Bank of America Visa fue de $1838 (*U.S. Airways Attaché Magazine*, diciembre de 2003). En una muestra de cargos mensuales a tarjetas de crédito los datos obtenidos son los siguientes.



236	1710	1351	825	7450
316	4135	1333	1584	387
991	3396	170	1428	1688

a. Calcule la media y la mediana.
b. Calcule el primero y tercer cuartil.
c. Calcule el rango y el rango intercuartílico.
d. Calcule la varianza y la desviación estándar.
e. El sesgo en este conjunto de datos es 2.12. Comente la forma de la distribución. ¿Esta es la forma que esperaría? ¿Por qué sí o por qué no?
f. ¿Hay observaciones atípicas en estos datos?

59. La oficina de censos de Estados Unidos proporciona estadísticas sobre las familias en ese país, informaciones como edad al contraer el primer matrimonio, estado civil actual y tamaño de la casa (www.census.gov, 20 de marzo de 2006). Los datos siguientes son edades al contraer el primer matrimonio en una muestra de hombres y en una muestra de mujeres.



Hombres	26	23	28	25	27	30	26	35	28
	21	24	27	29	30	27	32	27	25
Mujeres	20	28	23	30	24	29	26	25	
	22	22	25	23	27	26	19		

a. Determine la mediana en la edad de hombres y mujeres al contraer el primer matrimonio.
b. Calcule el primer y tercer cuartil tanto en los hombres como en las mujeres.
c. Hace 30 años la mediana en la edad al contraer el primer matrimonio era 25 años entre los hombres y 22 años entre las mujeres. ¿Qué indica esta información acerca de la edad a la que deciden contraer matrimonio los jóvenes de hoy en día?

60. El rendimiento de los dividendos son los beneficios anuales que paga una empresa por acción dividido entre el precio corriente en el mercado expresado como porcentaje. En una muestra de 10 empresas, los dividendos son los siguientes (*The Wall Street Journal*, 16 de enero de 2004).

Empresa	Porcentaje de rendimiento	Empresa	Porcentaje de rendimiento
Altria Group	5.0	General Motors	3.7
American Express	0.8	JPMorgan Chase	3.5
Caterpillar	1.8	McDonald's	1.6
Eastman Kodak	1.9	United Technology	1.5
ExxonMobil	2.5	Wal-Mart Stores	0.7

a. ¿Cuáles son la media y mediana de los rendimientos de dividendos?
b. ¿Cuál es la varianza y la desviación estándar?
c. ¿Qué empresa proporciona el mayor rendimiento de dividendos?
d. ¿Cuál es el punto z correspondiente a McDonalds? Interprete este punto z.
e. ¿Cuál es el punto z de General Motors? Interprete este punto z.
f. De acuerdo con los puntos z, ¿Hay algún dato atípico en la muestra?

61. El departamento de educación de Estados Unidos informa que cerca de 50% de los estudiantes universitarios toma un préstamo estudiantil como ayuda para cubrir sus gastos (Natural Center for Educational Studies, enero de 2006). Se tomó una muestra de los estudiantes que terminaron sus carreras teniendo una deuda sobre el préstamo estudiantil. Los datos muestran el monto en dólares de estas deudas:

10.1	14.8	5.0	10.2	12.4	12.2	2.0	11.5	17.8	4.0

a. Entre los estudiantes que toman un préstamo estudiantil, ¿cuál es la mediana en la deuda que tienen una vez terminados sus estudios?
b. ¿Cuál es la varianza y cuál la desviación estándar?

62. Los propietarios de negocios pequeños suelen contratar a empresas con servicio de nómina para que se encarguen del pago de sus empleados. Las razones son que encuentran regulaciones complicadas para el pago de impuestos y que las multas por errores en los impuestos de los empleados son elevadas. De acuerdo con el Internal Revenue Service, 26% de las declaraciones de impuestos de los empleados contienen errores que ocasionan multas a los dueños. (*The Wall Street Journal*, 30 de enero de 2006). La siguiente es una muestra de 20 multas a propietarios de negocios pequeños.





820	270	450	1010	890	700	1350	350	300	1200
390	730	2040	230	640	350	420	270	370	620

a. ¿Cuál es la media en multas?
b. ¿Cuál es la desviación estándar?
c. ¿Es una observación atípica la multa más alta, $2040?
d. ¿Cuáles son algunas de las ventajas que tienen los propietarios de los negocios pequeños al contratar una empresa de servicio de pago de nomina para que se ocupen del pago a sus empleados, incluyendo la declaración de impuestos de los empleados?

63. El transporte público y el automóvil son los dos medios que usa un empleado para ir a su trabajo cada día. Se presenta una muestra del tiempo requerido con cada medio. Los tiempos se dan en minutos.

Transporte público:	28	29	32	37	33	25	29	32	41	34
Automóvil:	29	31	33	32	34	30	31	32	35	33

a. Calcule la media muestral en el tiempo que se necesita con cada transporte.
b. Calcule la desviación estándar para cada transporte.
c. De acuerdo con los resultados en los incisos a y b ¿cuál será el medio de transporte preferido? Explique.
d. Para cada medio de transporte elabore un diagrama de caja. ¿Se confirma la conclusión que dio en el inciso c mediante una comparación de los diagramas de caja?

64. La National Association of Realtors informa sobre la mediana en el precio de una casa en Estados Unidos y sobre el aumento de esta mediana en los últimos cinco años. Use la muestra de precios de casas para responder a las preguntas siguientes.



995.9	48.8	175.0	263.5	298.0	218.9	209.0
628.3	111.0	212.9	92.6	2325.0	958.0	212.5

a. ¿Cuál es la mediana muestral de los precios de las casas?
b. En enero del 2001 la National Association of Realtors informó que la mediana en el precio de una casa en Estados Unidos era $139 300. ¿Cuál ha sido el incremento porcentual de la mediana en el precio de una casa en cinco años?
c. ¿Cuáles son el primer y tercer cuartiles de los datos muestrales?
d. Dé el resumen de cinco números para los precios de las casas.
e. ¿Existe alguna observación atípica en los datos?
f. ¿En la muestra cuál es la media en el precio de una casa? ¿Por qué prefiere la National Association of Realtors usar en sus informes la mediana en el precio de las casas?

65. Los datos siguientes son los gastos en publicidad (en millones de dólares) y los envíos en millones de barriles (bbls.) de las 10 principales marcas de cerveza.



Beer

Marca	Gastos en publicidad (millones de dólares)	Despachos en bbls (millones)
Budweiser	120.0	36.3
Bud Light	68.7	20.7
Miller Lite	100.1	15.9
Coors Light	76.6	13.2
Busch	8.7	8.1
Natural Light	0.1	7.1
Miller Genuine Draft	21.5	5.6
Miller High Life	1.4	4.4
Busch Lite	5.3	4.3
Milwaukee's Best	1.7	4.3

a. ¿Cuál es la covarianza muestral? ¿Indica que hay una relación positiva o negativa?
b. ¿Cuál es el coeficiente de correlación?

66. *Road &* Track proporciona la muestra siguiente de desgaste en llantas y la capacidad de carga máxima de llantas de automóviles.

Desgaste en llantas	Capacidad de carga máxima
75	853
82	1047
85	1135
87	1201
88	1235
91	1356
92	1389
93	1433
105	2039

a. Con estos datos elabore un diagrama de dispersión en el que el desgaste ocupe el eje *x*.
b. Calcule el coeficiente de correlación muestral. ¿Qué indica el coeficiente de correlación muestral acerca de la relación entre el desgaste y la capacidad de carga máxima?

67. Los datos siguientes presentan el seguimiento de la rentabilidad primaria por acción durante 52 semanas y los valores contables reportados por 10 empresas (*The Wall Street Journal*, 13 de marzo de 2000).

Empresa	Valor contable	Rentabilidad
Am Elec	25.21	2.69
Columbia En	23.20	3.01
Con Ed	25.19	3.13
Duke Energy	20.17	2.25
Edison Int'l	13.55	1.79
Enron Cp.	7.44	1.27
Peco	13.61	3.15
Pub Sv Ent	21.86	3.29
Southn Co.	8.77	1.86
Unicom	23.22	2.74

a. Elabore un diagrama de dispersión, que los valores contables ocupen el eje *x*.
b. Calcule el coeficiente de correlación muestral. ¿Qué indica este coeficiente acerca de la relación entre la rentabilidad por acción y el valor contable?

68. Una técnica de pronóstico conocida como promedios móviles emplea el promedio o la media de los *n* periodos más recientes para pronosticar el valor siguiente en los datos de una serie de tiempo. En un promedio móvil de tres periodos, se usan los datos de los tres periodos más recientes para calcular el pronóstico. Considere un producto que en los primeros tres meses de este año tuvo la demanda siguiente: enero (800 unidades), febrero (750 unidades) y marzo (900 unidades).
   a. ¿Cuál es pronóstico para abril empleando un promedio móvil de tres meses?
   b. A una variación de esta técnica se le conoce como promedios móviles ponderados. La ponderación permite que al calcular el pronóstico se le dé más importancia a los datos recientes de la serie de tiempo. Por ejemplo, en un promedio móvil de tres meses a los datos que tienen un mes de antigüedad se les da 3 como peso, 2 a los que tienen dos meses de antigüedad y 1 a los que tienen un mes. Con tales datos, calcule el pronóstico para abril usando promedios móviles de tres meses.

69. A continuación se presentan los días de plazo de vencimiento en una muestra de cinco fondos de mercado de dinero. Aparecen también las cantidades, en dólares, invertidas en los fondos. Emplee la media ponderada para determinar el número medio de días en los plazos de vencimiento de los dólares invertidos en estos cinco fondos de mercado de dinero.

Días de plazo de vencimiento	**Valor en dólares**
20	20
12	30
7	10
5	15
6	10

70. Un sistema de radar de la policía vigila los automóviles en una carretera que permite una velocidad máxima de 55 millas por hora. La siguiente es una distribución de frecuencias de las velocidades.

Velocidad (millas por hora)	**Frecuencia**
45–49	10
50–54	40
55–59	150
60–64	175
65–69	75
70–74	15
75–79	10
Total	475

   a. ¿Cuál es la velocidad media de los automóviles en esta carretera?
   b. Calcule la varianza y la desviación estándar.

Caso problema 1 Las tiendas Pelican

Las tiendas Pelican, una división de National Clothing, es una cadena de tiendas de ropa para mujer con sucursales por todo Estados Unidos. En fechas recientes la cadena realizó una promoción en la que envió cupones de descuento a clientes de otras tiendas de National Clothing. Los datos obtenidos en una muestra de 100 pagos con tarjeta de crédito en las tiendas Pelican, durante un día de la promoción, aparecen en el archivo titulado PelicanStores. En la tabla 3.14 se muestra parte de este conjunto de datos. El modo de pago Propietary card se refiere a pagos realizados con tarjeta de crédito de National Clothing. A los clientes que hicieron compras con un cupón de descuento se les denomina aquí promocionales y a quienes hicieron sus compras sin emplear cupón de descuento se les denomina regulares. Como a los clientes de las tiendas Pelican no se les enviaron cupones promocionales, los directivos consideran que las ventas hechas a las personas que presentaron un cupón de descuento son ventas que de otro modo no se hubieran realizado. Es obvio que Pelican espera que los clientes promocionales continúen comprando en sus tiendas.

La mayor parte de las variables que aparecen en la tabla 3.14 se explican por sí mismas, pero dos de ellas deben ser aclaradas.

Artículos	Número de artículos comprados
Ventas netas	Cantidad cargada a la tarjeta de crédito

Los directivos de Pelican desean emplear estos datos muestrales para tener información acerca de sus clientes y evaluar la promoción de los cupones de descuento.

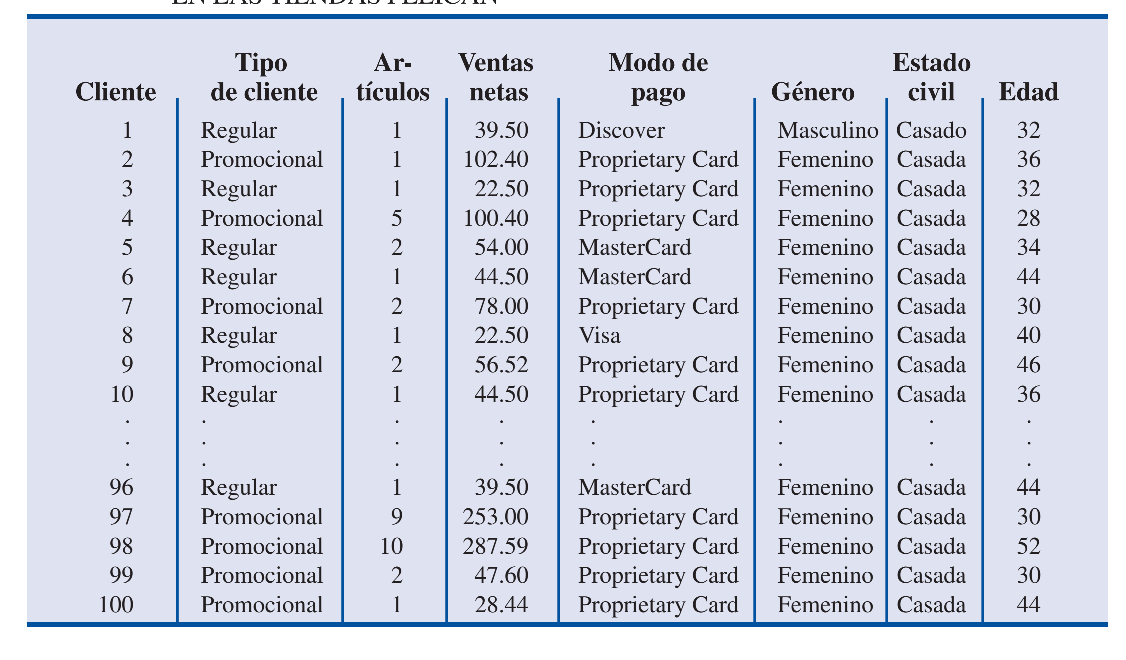
Informe para los directivos

Use los métodos de la estadística descriptiva presentados en este capítulo para resumir los datos y comente sus hallazgos. Su informe debe contener, por lo menos, lo siguiente:

1. Estadísticos descriptivos sobre las ventas netas y sobre las ventas a los distintos tipos de clientes.
2. Estadísticos descriptivos respecto de la relación entre edad y ventas netas.

TABLA 3.14 MUESTRA DE 100 COMPRAS CON TARJETA DE CRÉDITO REALIZADAS EN LAS TIENDAS PELICAN





Cliente	Tipo de cliente	Artículos	Ventas netas	Modo de pago	Género	Estado civil	Edad
1	Regular	1	39.50	Discover	Masculino	Casado	32
2	Promocional	1	102.40	Proprietary Card	Femenino	Casada	36
3	Regular	1	22.50	Proprietary Card	Femenino	Casada	32
4	Promocional	5	100.40	Proprietary Card	Femenino	Casada	28
5	Regular	2	54.00	MasterCard	Femenino	Casada	34
6	Regular	1	44.50	MasterCard	Femenino	Casada	44
7	Promocional	2	78.00	Proprietary Card	Femenino	Casada	30
8	Regular	1	22.50	Visa	Femenino	Casada	40
9	Promocional	2	56.52	Proprietary Card	Femenino	Casada	46
10	Regular	1	44.50	Proprietary Card	Femenino	Casada	36
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
96	Regular	1	39.50	MasterCard	Femenino	Casada	44
97	Promocional	9	253.00	Proprietary Card	Femenino	Casada	30
98	Promocional	10	287.59	Proprietary Card	Femenino	Casada	52
99	Promocional	2	47.60	Proprietary Card	Femenino	Casada	30
100	Promocional	1	28.44	Proprietary Card	Femenino	Casada	44

Caso problema 2 Industria cinematográfica

La industria cinematográfica es un negocio muy competido. En más de 50 estudios se producen 300 a 400 películas por año y el éxito financiero de estas películas varía en forma considerable. Las variables usuales para medir el éxito de una película son ventas brutas (en millones de dólares) en el fin de semana del estreno, ventas brutas totales (en millones de dólares), número de salas donde se presenta la película, semanas en las que la película se encuentra entre las 60 mejores en ventas brutas. Los datos de una muestra de 100 películas producidas en 2005 se encuentran en el archivo titulado Movies. La tabla 3.15 muestra los datos de las 10 primeras películas que se encuentran en este archivo.

Informe para los directivos

Use los métodos numéricos de la estadística descriptiva presentados en este capítulo para averiguar cómo contribuyen estas variables al éxito de una película. Su informe debe contener lo siguiente.

1. Estadísticos descriptivos para cada una de las cuatro variables con un análisis sobre la información que la estadística descriptiva proporciona acerca de la industria del cine.
2. ¿Hay alguna película que deba ser considerada como una observación atípica de alto desempeño?
3. Los estadísticos descriptivos muestran la relación entre ventas brutas y cada una de las otras variables. Argumente.

TABLA 3.15 DATOS DEL ÉXITO DE 10 PELÍCULAS



Película	Ventas brutas en el estreno (en millones de dólares)	Ventas brutas totales (en millones de dólares)	Número de salas	Semanas en las 60 mejores
Coach Carter	29.17	67.25	2574	16
Ladies in Lavender	0.15	6.65	119	22
Batman Begins	48.75	205.28	3858	18
Unleashed	10.90	24.47	1962	8
Pretty Persuasion	0.06	0.23	24	4
Fever Pitch	12.40	42.01	3275	14
Harry Potter and the Goblet of Fire	102.69	287.18	3858	13
Monster-in-Law	23.11	82.89	3424	16
White Noise	24.11	55.85	2279	7
Mr. and Mrs. Smith	50.34	186.22	3451	21

Caso problema 3 Las escuelas de negocios de Asia-Pacífico

En la actualidad se ha vuelto mundial el interés por tener un grado superior en estudios de negocios. En una investigación se encontró que en Asia cada vez más personas eligen una maestría en administración de negocios como camino hacia el éxito corporativo. De esta manera, en las escuelas de Asia–Pacífico, el número de solicitudes a cursos de maestría en administración de negocios sigue aumentando.



En esa región miles de personas suspenden sus carreras y pasan dos años en estudios para obtener una formación teórica en negocios. Los cursos en estas escuelas son bastante pesados y comprenden economía, banca, marketing, ciencias de la conducta, relaciones laborales, toma de decisiones, pensamiento estratégico, derecho internacional en negocios y otras áreas. En los datos que se presentan en la tabla 3.16 aparecen algunas de las características de las principales escuelas de negocios de Asia–Pacífico.

TABLA 3.16 DATOS DE 25 ESCUELAS DE NEGOCIOS EN ASIA–PACÍFICO

Escuela de negocios	Estudiantes de tiempo completo	Estudiantes por facultad	Colegiatura para estudiantes locales ($)	Colegiatura para estudiantes de fuera ($)	Edad	% de extranjeros	GMAT	Examen de inglés	Experiencia laboral	Salario inicial ($)
Melbourne Business School	200	5	24 420	29 600	28	47	Sí	No	Sí	71 400
University of New South Wales (Sydney)	228	4	19 993	32 582	29	28	Sí	No	Sí	65 200
Indian Institute of Management (Ahmedabad)	392	5	4 300	4 300	22	0	No	No	No	7 100
Chinese University of Hong Kong	90	5	11 140	11 140	29	10	Sí	No	No	31 000
International University of Japan (Niigata)	126	4	33 060	33 060	28	60	Sí	Sí	No	87 000
Asian Institute of Management (Manila)	389	5	7 562	9 000	25	50	Sí	No	Sí	22 800
Indian Institute of Management (Bangalore)	380	5	3 935	16 000	23	1	Sí	No	No	7 500
National University of Singapore	147	6	6 146	7 170	29	51	Sí	Sí	Sí	43 300
Indian Institute of Management (Calcutta)	463	8	2 880	16 000	23	0	No	No	No	7 400
Australian National University (Canberra)	42	2	20 300	20 300	30	80	Sí	Sí	Sí	46 600
Nanyang Technological University (Singapore)	50	5	8 500	8 500	32	20	Sí	No	Sí	49 300
University of Queensland (Brisbane)	138	17	16 000	22 800	32	26	No	No	Sí	49 600
Hong Kong University of Science and Technology	60	2	11 513	11 513	26	37	Sí	No	Sí	34 000
Macquarie Graduate School of Management (Sydney)	12	8	17 172	19 778	34	27	No	No	Sí	60 100
Chulalongkorn University (Bangkok)	200	7	17 355	17 355	25	6	Sí	No	Sí	17 600
Monash Mt. Eliza Business School (Melbourne)	350	13	16 200	22 500	30	30	Sí	Sí	Sí	52 500
Asian Institute of Management (Bangkok)	300	10	18 200	18 200	29	90	No	Sí	Sí	25 000
University of Adelaide	20	19	16 426	23 100	30	10	No	No	Sí	66 000
Massey University (Palmerston North, New Zealand)	30	15	13 106	21 625	37	35	No	Sí	Sí	41 400
Royal Melbourne Institute of Technology Business Graduate School	30	7	13 880	17 765	32	30	No	Sí	Sí	48 900
Jamnalal Bajaj Institute of Management Studies (Bombay)	240	9	1 000	1 000	24	0	No	No	Sí	7 000
Curtin Institute of Technology (Perth)	98	15	9 475	19 097	29	43	Sí	No	Sí	55 000
Lahore University of Management Sciences	70	14	11 250	26 300	23	2.5	No	No	No	7 500
Universiti Sains Malaysia (Penang)	30	5	2 260	2 260	32	15	No	Sí	Sí	16 000
De La Salle University (Manila)	44	17	3 300	3 600	28	3.5	Sí	No	Sí	13 100

Informe para los directivos

Use los métodos de la estadística descriptiva para resumir los datos de la tabla 3.16. Argumente sobre sus hallazgos.

1. Para cada variable presente un resumen del conjunto de datos. Haga comentarios e interpretaciones con base en máximos y mínimos, así como en las medias y proporciones adecuadas. ¿Qué conclusiones nuevas proporcionan estos estadísticos descriptivos respecto de las escuelas de negocios de Asia–Pacífico?
2. Resuma los datos para hacer las comparaciones siguientes:
   a. Diferencias entre las colegiaturas para alumnos locales y de fuera.
   b. Diferencias entre los salarios promedio iniciales para egresados de escuelas que exigen experiencia laboral y de escuelas que no la exigen.
   c. Discrepancias entre los salarios promedio iniciales de egresados de escuelas que exigen una prueba de inglés y de escuelas que no la exigen.
3. ¿Parece haber relación entre los salarios iniciales y las colegiaturas?
4. Presente cualquier gráfica y resumen numérico que pueda servir para comunicar a otras personas la información presentada en la tabla 3.16.

Apéndice 3.1 Estadística descriptiva usando Minitab

En este apéndice se describe cómo usar Minitab para obtener estadísticos descriptivos. En la tabla 3.1 aparecen los sueldos iniciales de 12 recién egresados de la carrera de administración. En el panel A de la figura 3.11 están los estadísticos descriptivos obtenidos para resumir los datos usando Minitab. A continuación se dan las definiciones de los títulos que se observan en el panel A.

N	número de valores en los datos
N*	número de datos faltantes
Mean	media
SE Mean	error estándar de la media
StDev	desviación estándar
Minimum	valor mínimo (menor) en los datos
Q1	primer cuartil
Median	mediana
Q3	tercer cuartil
Maximum	valor máximo (mayor) en los datos

El título SE mean se refiere al *error estándar de la media.* Este valor se obtiene dividiendo la desviación estándar entre la raíz cuadrada de N. La interpretación y uso de esta medición se verá en el capítulo 7, cuando se introduzca el tema del muestreo y de la distribución muestral.

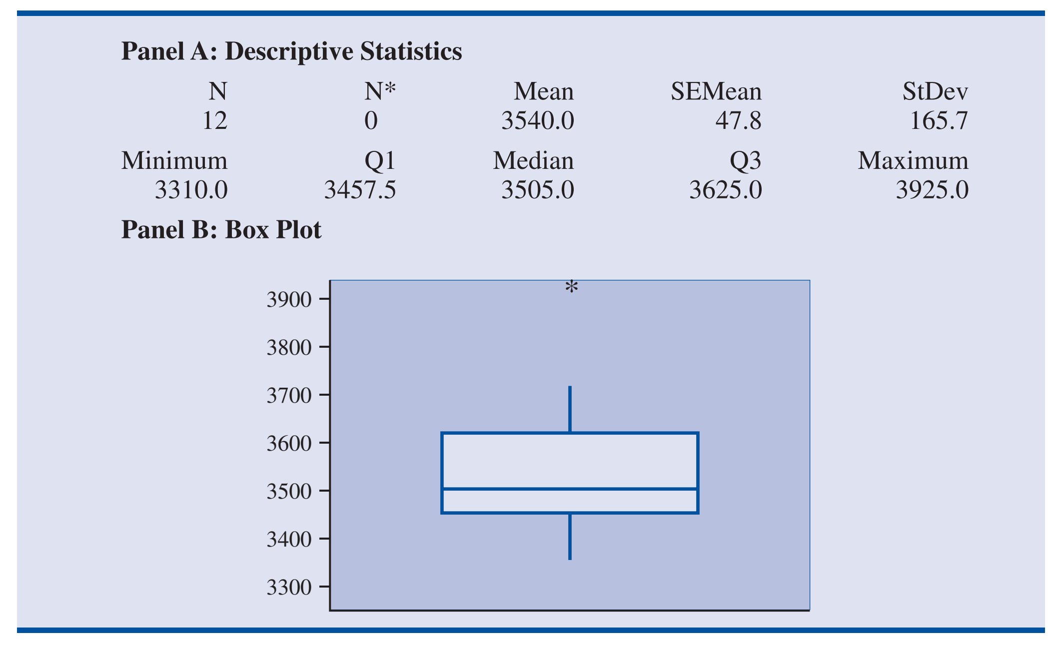
Aunque en los resultados de Minitab no aparecen el rango, el rango intercuartílico, la varianza y el coeficiente de variación, estas medidas son fáciles de calcular a partir de los resultados que aparecen en la figura 3.11; se calculan como sigue.

$$\text{Rango} = \text{Máximo} - \text{Mínimo}$$
$$\text{RIC} = Q_3 - Q_1$$
$$\text{Varianza} = (\text{StDev})^2$$
$$\text{Coeficiente de variación} = (\text{StDev/Media}) \times 100$$

Por último, observe que los cuartiles que da Minitab, $Q_1 = 3457.5$ y $Q_3 = 3625$, son ligeramente diferentes a los calculados en la sección 3.1. Esto se debe al empleo de convenciones* di-

*Cuando se tienen n observaciones ordenadas de menor a mayor (en orden ascendente), para localizar los cuartiles Q_1 y Q_3 Minitab usa las posiciones dadas por $(n + 1)/4$ y $3(n + 1)/4$, respectivamente. Si se obtiene un número fraccionario, Minitab interpola entre los valores de los datos adyacentes ordenados para determinar el cuartil correspondiente.

FIGURA 3.11 ESTADÍSTICOS DESCRIPTIVOS Y DIAGRAMA DE CAJA PROPORCIONADOS POR MINITAB



Panel A: Descriptive Statistics

N	N*	Mean	SEMean	StDev
12	0	3540.0	47.8	165.7

Minimum	Q1	Median	Q3	Maximum
3310.0	3457.5	3505.0	3625.0	3925.0

Panel B: Box Plot

ferentes para identificar los cuartiles. De manera que los valores Q_1 y Q_3 obtenidos con una convención quizá no sean idénticos a los valores Q_1 y Q_3 obtenidos con otra. Sin embargo, estas diferencias tienden a ser despreciables y los resultados no afectan al hacer las interpretaciones relacionadas con los cuartiles.

Ahora verá cómo se generan los estadísticos que aparecen en la figura 3.11. Los datos de los sueldos iniciales se encuentran en al columna C2 de la hoja de cálculo de Minitab. Para generar los estadísticos descriptivos realice los pasos siguientes:



Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Basic Statistics**
Paso 3. Elegir **Display Descriptive Statistics**
Paso 4. Cuando aparece el cuadro de diálogo Display Descriptive Statistics:
Ingresar C2 en el cuadro **Variables**
Dar clic en **OK**

El panel B de la figura 3.11 es un diagrama de caja obtenido con Minitab y contiene entre el primer y tercer cuartil 50% de los datos. La línea dentro de la caja corresponde a la mediana. El asterisco indica que hay una observación atípica en 3925.

Con los pasos siguientes se genera el diagrama de caja que aparece en la figura 3.11.

Paso 1. Seleccionar el menú **Graph**
Paso 2. Elegir **Boxplot**
Paso 3. Elegir **Simple** y hacer clic en **OK**
Paso 4. Cuando aparezca el cuadro de diálogo Boxplot-One Y, Simple:
Ingresar C2 en el cuadro **Graph variables**
Hacer clic en **OK**

La medida del sesgo tampoco aparece como parte de los resultados estándar de estadística descriptiva que proporciona Minitab. Sin embargo, puede incluirse mediante los pasos siguientes.

FIGURA 3.12 COVARIANZA Y CORRELACIÓN OBTENIDAS USANDO MINITAB CON LOS DATOS DEL NÚMERO DE COMERCIALES Y VENTAS

```
Covariances: No. of Commercials, Sales Volume

              No. of Comme  Sales Volume
No. of Comme       2.22222
Sales Volume      11.00000      62.88889

Correlations: No. of Commercials, Sales Volume

Pearson correlation of No. of Commercials and Sales Volume = 0.930
P-Value = 0.000
```

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Basic Statistics**
Paso 3. Elegir **Display Descriptive Statistics**
Paso 4. Cuando aparezca el cuadro de diálogo Display Descriptive Statistics:
Clic en **Statistics**
Elegir **Skewness**
Clic en **OK**
Clic en **OK**

La medida del sesgo, 1.09, aparecerá en su hoja de cálculo.



Stereo

La figura 3.12 muestra los resultados que da Minitab para la covarianza y la correlación con los datos de la tienda de equipos de sonido presentados en la tabla 3.7. En la parte de la figura que corresponde a la covarianza, *No. of Comme* denota el número de semanas que se televisaron los comerciales y *Sales Volume* las ventas durante la semana siguiente. El valor que aparece en la columna *No. of Comme* y en el renglón *Sales Volume*, 11, es la covarianza muestral que se calculó en la sección 3.5. El valor de la columna *No. of Comme* y en el renglón No. of Comme, 2.22222, es la varianza muestral del número de comerciales, y el valor que se encuentra en la columna *Sales Volume* y en el renglón *Sales Volume*, 62. 88889, es la varianza muestral de las ventas. El coeficiente de correlación muestral, 0.930, aparece en los resultados, en la parte correspondiente a la correlación. Nota: la interpretación del valor-$p = 0.000$ se verá en el capítulo 9.

Ahora se describe cómo obtener la información que se muestra en la figura 3.12. En la columna C2 de la hoja de cálculo de Minitab ingrese los datos del número de comerciales y en la columna C3 los datos de las ventas. Los pasos necesarios para obtener los resultados que se muestran en los tres primeros renglones de la figura 3.12 son los siguientes.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Basic Statistics**
Paso 3. Elegir **Covariance**
Paso 4. Cuando aparezca el cuadro de diálogo Covariance:
Ingresar C2 C3 en el cuadro **Variable**
Clic en **OK**

Para obtener el resultado correspondiente a la correlación, que se observa en la tabla 3.12, sólo hay que hacer una modificación a estos pasos para la covarianza. En el paso 3 seleccione la opción **Correlation**.

Apéndice 3.2 Estadísticos descriptivos usando Excel

Emplee Excel para generar los estadísticos descriptivos vistos en este capítulo. Ahora aprenderá a usar Excel para generar diversas medidas de localización y de variabilidad para una variable, así como la covarianza y el coeficiente de correlación para medir la asociación entre dos variables.

FIGURA 3.13 USO DE LAS FUNCIONES DE EXCEL PARA CALCULAR LA MEDIA, MEDIANA, MODA, VARIANZA Y DESVIACIÓN ESTÁNDAR

	A	B	C	D	E	F
1	Graduate	Starting Salary		Mean	=AVERAGE(B2:B13)	
2	1	3450		Median	=MEDIAN(B2:B13)	
3	2	3550		Mode	=MODE(B2:B13)	
4	3	3650		Variance	=VAR(B2:B13)	
5	4	3480		Standard Deviation	=STDEV(B2:B13)	
6	5	3355				
7	6	3310				
8	7	3490				
9	8	3730				
10	9	3540				
11	10	3925				
12	11	3520				
13	12	3480				
14						

	A	B	C	D	E	F
1	Graduate	Starting Salary		Mean	3540	
2	1	3450		Median	3505	
3	2	3550		Mode	3480	
4	3	3650		Variance	27440.91	
5	4	3480		Standard Deviation	165.65	
6	5	3355				
7	6	3310				
8	7	3490				
9	8	3730				
10	9	3540				
11	10	3925				
12	11	3520				
13	12	3480				
14						

Uso de las funciones de Excel



StartSalary

Excel tiene funciones para calcular media, mediana, moda, varianza muestral y desviación estándar muestral. Con los datos de los sueldos iniciales de la tabla 3.1 ilustrará el uso de las funciones de Excel para calcular la media, mediana, moda, varianza muestral y desviación estándar muestral. Al ir siguiendo los pasos necesarios, consulte la figura 3.13. Ingrese los datos en la columna B.

Para calcular la media emplee la función AVERAGE (PROMEDIO) de Excel ingresando la fórmula siguiente en la celda E1:

=AVERAGE(B2:B13)

De manera similar ingrese en las celdas E2:E5 las fórmulas =MEDIANA(B2:B13), =MODA(B2:B13), =VAR(B2:B13) y =DESVEST(B2:B13) para calcular, respectivamente, la mediana, moda, varianza y desviación estándar. La hoja de cálculo que aparece en primer plano muestra que los valores calculados usando las funciones de Excel son iguales a los ya calculados en este capítulo.

Excel tiene también funciones para calcular la covarianza y el coeficiente de correlación. Al usar estas funciones debe tener cuidado, dado que la función covarianza trata a los datos como población y la función correlación como muestra. Por tanto, los resultados obtenidos con la función covarianza de Excel deben ajustarse para obtener la covarianza muestral. Se le muestra cómo usar estas funciones de Excel para el cálculo de la covarianza muestral y del coeficiente de correlación muestral empleando los datos de la tienda que vende equipos de sonido y que se presentaron en la figura 3.14.



Stereo

FIGURA 3.14 USO DE LAS FUNCIONES DE EXCEL PARA CALCULAR LA COVARIANZA Y LA CORRELACIÓN

	A	B	C	D	E	F	G
1	Week	Commercials	Sales		Population Covariance	=COVAR(B2:B11:C2:C11)	
2	1	2	50		Sample Correlation	=CORREL(B2:B11,C2:C11)	
3	2	5	57				
4	3	1	41				
5	4	3	54				
6	5	4	54				
7	6	1	38				
8	7	5	63				
9	8	3	48				
10	9	4	59				
11	10	2	46				
12							

	A	B	C	D	E	F	G
1	Week	Commercials	Sales		Population Covariance	9.90	
2	1	2	50		Sample Correlation	0.93	
3	2	5	57				
4	3	1	41				
5	4	3	54				
6	5	4	54				
7	6	1	38				
8	7	5	63				
9	8	3	48				
10	9	4	59				
11	10	2	46				
12							

La función covarianza de Excel, COVAR, se emplea para calcular la covarianza poblacional ingresando la fórmula siguiente en la celda F1

=COVAR(B2:B11,C2:C11)

De manera similar ingrese la fórmula: CORREL(B2:B11,C2:C11) para calcular el coeficiente de correlación muestral. En la hoja de cálculo que aparece en primer plano aparecen los valores obtenidos usando estas funciones de Excel. Observe que el valor del coeficiente de correlación muestral (0.93) es el mismo que obtuvo empleando la ecuación (3.12). Sin embargo, el resultado obtenido, 9.9, mediante la función COVAR de Excel, lo obtuvo tratando los datos como población. Por tanto, es necesario ajustar este resultado de Excel para obtener la covarianza muestral. Este ajuste es bastante sencillo. En primer lugar hay que observar que en la fórmula para la covarianza poblacional, ecuación (3.11), requiere dividir entre el número total de observaciones en el conjunto de datos. En cambio, en la fórmula para la covarianza muestral, ecuación (3.10), requiere dividir entre el número total de observaciones menos 1. Entonces, para usar este resultado de Excel, 9.9, para calcular la covarianza muestral, simplemente multiplique 9.9 por $n/(n - 1)$. Como $n = 10$, se tiene

$$s_{xy} = \left(\frac{10}{9}\right)9.9 = 11$$

De esta manera la covarianza muestral de los datos de la tienda de equipos para sonido es 11.

Uso de las herramientas de Excel para estadísticos descriptivos

Como se mostró, Excel tiene funciones estadísticas que permiten calcular los estadísticos descriptivos de un conjunto de datos. Estas funciones sirven para calcular dichos estadísticos de uno en uno (por ejemplo, la media, la varianza, etc.). Excel cuenta también con diversas herramientas para el análisis de datos. Una de estas herramientas llamada Estadística descriptiva, permite calcular varios estadísticos descriptivos de una sola vez. A continuación se le muestra cómo usar

FIGURA 3.15 USO DE LAS HERRAMIENTAS DE EXCEL PARA ESTADÍSTICOS DESCRIPTIVOS

	A	B	C	D	E	F
1	Graduate	Starting Salary		*Starting Salary*		
2	1	3450				
3	2	3550		**Mean**	3540	
4	3	3650		Standard Error	47.82	
5	4	3480		**Median**	3505	
6	5	3355		**Mode**	3480	
7	6	3310		**Standard Deviation**	165.65	
8	7	3490		**Sample Variance**	27440.91	
9	8	3730		Kurtosis	1.7189	
10	9	3540		**Skewness**	1.0911	
11	10	3925		**Range**	615	
12	11	3520		**Minimum**	3310	
13	12	3480		**Maximum**	3925	
14				**Sum**	42480	
15				**Count**	12	
16						



StartSalary

esta herramienta para calcular los estadísticos descriptivos del conjunto de datos referidos a los sueldos iniciales presentados en la tabla 3.1. Consulte la figura 3.15 a medida que se le describen los pasos necesarios.

Paso 1. Seleccionar el menú **Herramientas**
Paso 2. Elegir **Análisis de datos**
Paso 3. Cuando aparezca el cuadro de diálogo Análisis de datos:
Elegir **Estadística descriptiva**
Clic en **OK**
Paso 4. Cuando aparezca el cuadro de diálogo Estadística descriptiva:
Ingresar B1:B13 en el cuadro **Rango de entrada**
Seleccionar **Agrupados por Columnas**
Seleccionar **Rótulos en la primera fila**
Seleccionar **Rango de salida**
Ingresar D1 en la caja para el rango de salida (para identificar la esquina superior izquierda de la hoja de cálculo en la que aparecerá la estadística descriptiva)
Seleccionar **Resumen de estadísticas**
Clic en **OK.**

Las celdas D1:D15 de la figura 3.15 muestran la estadística descriptiva obtenida con Excel. Las entradas en negritas son los estadísticos descriptivos que se estudiaron en este capítulo. Los estadísticos descriptivos que no están en negritas se estudiarán en capítulos subsiguientes o en textos más avanzados.

CAPÍTULO 4

Introducción a la probabilidad

CONTENIDO



LA ESTADÍSTICA *en* LA PRÁCTICA

LA EMPRESA ROHM AND HASS*

Filadelfia, Pensilvania

Rohm and Hass es el principal productor de materiales especiales, entre los que se encuentran materiales electrónicos, polímeros para pinturas y artículos para el cuidado personal. Los productos de esta empresa permiten la creación de bienes de consumo de vanguardia en mercados como el farmacéutico, el de alimentos, el de suministros para la construcción, equipos de comunicación y productos para el hogar. La fuerza de trabajo de la empresa es de más de 17 000 personas y sus ventas anuales son de $8 mil millones. Una red de más de 100 puntos de fabricación, investigación técnica y servicio al cliente proporciona los productos y servicios de Rohm and Hass en 27 países.

En el área de productos químicos especiales, la empresa ofrece diversos productos químicos destinados a satisfacer las especificaciones únicas de sus clientes. Para un cliente determinado, la empresa produce un catalizador caro que el cliente emplea en sus procesos químicos. Algunos, pero no todos los lotes que produce la empresa satisfacen las especificaciones del producto. El contrato estipula que el cliente debe probar cada lote después de recibirlo y determinar si el catalizador podrá realizar la función esperada. Los lotes que no pasen la prueba del cliente serán regresados. Con el tiempo, la experiencia ha mostrado que el cliente acepta 60% de los lotes y regresa 40%. Ni el cliente ni la empresa estaban satisfechos con este servicio.

La empresa examinó la posibilidad de, antes de enviar el lote, replicar la prueba que hacía el cliente. Sin embargo, los elevados costos del equipo especial que se necesitaba para la prueba hicieron que esta posibilidad no fuera factible. Los químicos de la empresa encargados del problema propusieron una prueba diferente de costo bajo que se podía practicar antes de enviar el lote al cliente. La empresa creyó que la nueva prueba podría indicar si el catalizador pasaría la compleja prueba que practicaba el cliente.

*Los autores agradecen a Michael Haskell, de la subsidiaria Morton International de Rohm and Hass por haberles proporcionado este artículo para *La estadística en la práctica.*



Una nueva prueba antes de enviar el lote al cliente mejora el servicio al cliente. © Keith Word/Stone.

La pregunta era: ¿cuál es la probabilidad de que el catalizador pase la prueba del cliente dado que pasó la nueva prueba antes de enviar el lote?

La empresa produjo una muestra del catalizador y la sometió a la nueva prueba. Entonces sólo los lotes de catalizador que pasaban la prueba se enviaban al cliente. Mediante el análisis de probabilidad de los datos se supo que si el catalizador pasaba la nueva prueba antes de ser enviado al cliente, la probabilidad de que el catalizador pasara la prueba del cliente era 0.909. O que si el catalizador pasaba la prueba de la empresa, la probabilidad de que no pasara la prueba del cliente y fuera rechazado era 0.091. El análisis de probabilidad aportó evidencias para poner en uso el procedimiento de la prueba antes de enviar el lote. Esta nueva prueba tuvo una mejora inmediata en el servicio al cliente y redujo tanto los costos como los gastos de envío y el manejo de los lotes regresados.

A la probabilidad de que un lote sea aceptado por el cliente, dado que pasó la nueva prueba, se le llama probabilidad condicional. En este capítulo aprenderá cómo calcular la probabilidad condicional y otras probabilidades útiles en la toma de decisiones.

Los administradores sustentan sus decisiones en un análisis de incertidumbres como las siguientes:

1. ¿Qué posibilidades hay de que disminuyan las ventas si aumentamos los precios?
2. ¿Qué posibilidad hay de que un método nuevo de ensamblado aumente la productividad?
3. ¿Cuáles son las posibilidades de que el producto se tenga listo a tiempo?
4. ¿Qué oportunidad existe de que una nueva invención sea rentable?

Algunos de los primeros trabajos sobre probabilidad se dieron en una serie de cartas entre Pierre de Fermat y Blaise Pascal durante el año de 1650.

La **probabilidad** es una medida numérica de la posibilidad de que ocurra un evento. Por tanto, las probabilidades son una medida del grado de incertidumbre asociado con cada uno de los eventos previamente enunciados. Si cuenta con las probabilidades, tiene la capacidad de determinar la posibilidad de ocurrencia que tiene cada evento.

Los valores de probabilidad se encuentran en una escala de 0 a 1. Los valores cercanos a 0 indican que las posibilidades de que ocurra un evento son muy pocas. Los cercanos a 1 indican que es casi seguro que ocurra un evento. Otras probabilidades entre cero y uno representan distintos grados de posibilidad de que ocurra un evento. Por ejemplo, si considera el evento "que llueva mañana", se entiende que si el pronóstico del tiempo dice "la probabilidad de que llueva es cercana a cero", implica que casi no hay posibilidades de que llueva. En cambio, si informan que la probabilidad de que llueva es 0.90, sabe que es muy posible que llueva. La probabilidad de 0.50 indica que es igual de posible que llueva como que no llueva. En la figura 4.1 se presenta la probabilidad como una medida numérica de la posibilidad de que ocurra un evento.



4.1 Experimentos, reglas de conteo y asignación de probabilidades

En el contexto de la probabilidad, un **experimento** es definido como un proceso que genera resultados definidos. Y en cada una de las repeticiones del experimento, habrá uno y sólo uno de los posibles resultados experimentales. A continuación se dan varios ejemplos de experimentos con sus correspondientes resultados.

Experimento	Resultado experimental
Lanzar una moneda	Cara, cruz
Tomar una pieza para inspeccionarla	Con defecto, sin defecto
Realizar una llamada de ventas	Hay compra, no hay compra
Lanzar un dado	1, 2, 3, 4, 5, 6
Jugar un partido de futbol	Ganar, perder, empatar

Al especificar todos los resultados experimentales posibles, está definiendo el **espacio muestral** de un experimento.

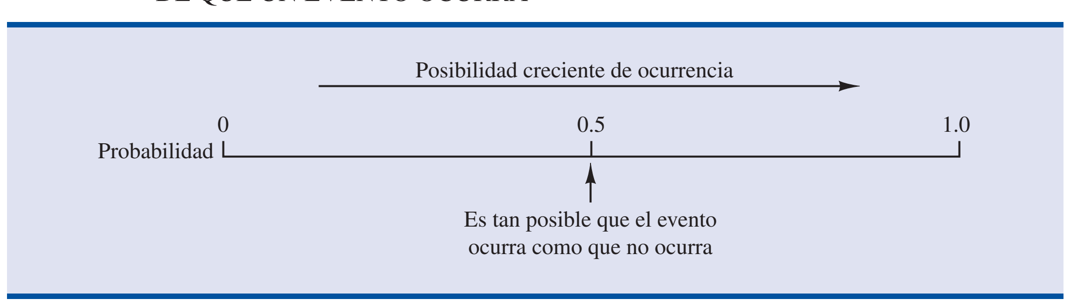
ESPACIO MUESTRAL

El espacio muestral de un experimento es el conjunto de todos los resultados experimentales.

A los resultados experimentales también se les llama puntos muestrales.

A un resultado experimental también se le llama **punto muestral** para identificarlo como un elemento del espacio muestral.

FIGURA 4.1 PROBABILIDAD COMO MEDIDA NUMÉRICA DE LA POSIBILIDAD DE QUE UN EVENTO OCURRA



Considere el primer experimento presentado en la tabla anterior, lanzar una moneda. La cara de la moneda que caiga hacia arriba —cara o cruz— determina el resultado experimental (puntos muestrales). Si denota con S el espacio muestral, puede emplear la notación siguiente para describir el espacio muestral.

$$S = \{\text{Cara, cruz}\}$$

En el segundo experimento de la tabla –tomar una pieza para revisarla– puede describir el espacio muestral como sigue:

$$S = \{\text{Defectuosa, no defectuosa}\}$$

Los dos experimentos descritos tienen dos resultados experimentales (puntos muestrales). Pero, observe ahora el cuarto experimento enumerado en la tabla, lanzar un dado. Los resultados experimentales, definidos por el número de puntos del dado en la cara que cae hacia arriba, son los seis puntos del espacio muestral de este experimento.

$$S = \{1, 2, 3, 4, 5, 6\}$$

Reglas de conteo, combinaciones y permutaciones

Al asignar probabilidades es necesario saber identificar y contar los resultados experimentales. A continuación tres reglas de conteo que son muy utilizadas.

Experimentos de pasos múltiples La primera regla de conteo sirve para experimentos de pasos múltiples. Considere un experimento que consiste en lanzar dos monedas. Defina los resultados experimentales en términos de las caras y cruces que se observan en las dos monedas. ¿Cuántos resultados experimentales tiene este experimento? El experimento de lanzar dos monedas es un experimento de dos pasos: el paso 1 es lanzar la primera moneda y el paso 2 es lanzar la segunda moneda. Si se emplea H para denotar cara y T para denotar cruz, (H, H) será el resultado experimental en el que se tiene cara en la primera moneda y cara en la segunda moneda. Si continúa con esta notación, el espacio muestral (S) en este experimento del lanzamiento de monedas será el siguiente:

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

Por tanto, hay cuatro resultados experimentales. En este caso es fácil enumerar todos los resultados experimentales.

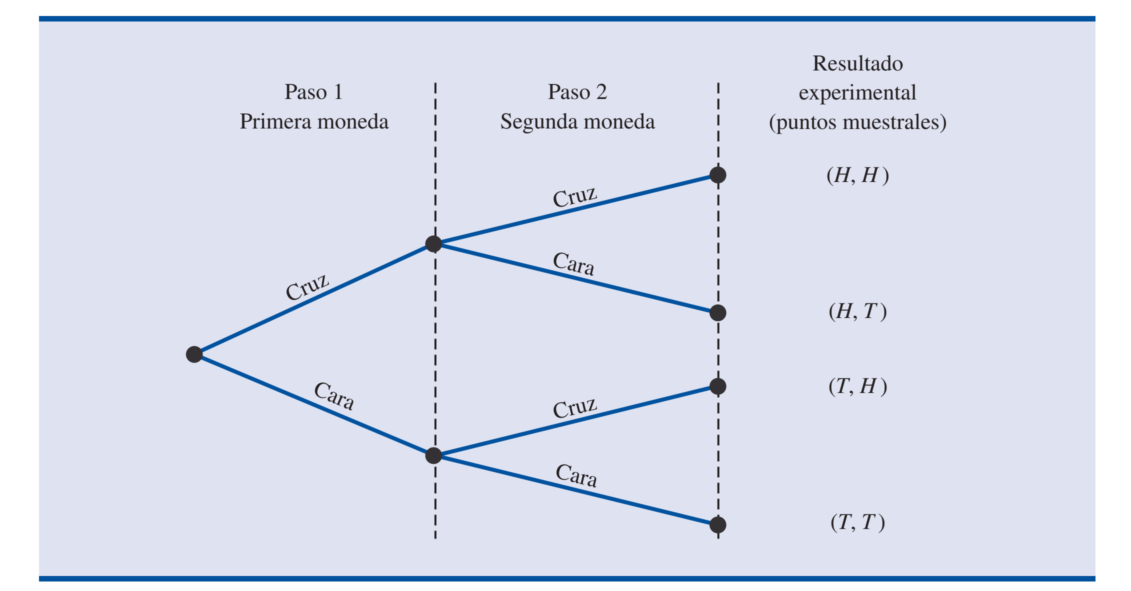
La regla de conteo para experimentos de pasos múltiples permite determinar el número de resultados experimentales sin tener que enumerarlos.

REGLA DE CONTEO PARA EXPERIMENTOS DE PASOS MÚLTIPLES

Un experimento se describe como una sucesión de k pasos en los que hay n_1 resultados posibles en el primer paso, n_2 resultados posibles en el segundo paso y así en lo sucesivo, entonces el número total de resultados experimentales es $(n_1)\,(n_2)\,.\,.\,.\,(n_k)$.

Si considera el experimento del lanzamiento de dos monedas como la sucesión de lanzar primero una moneda ($n_1 = 2$) y después lanzar la otra ($n_2 = 2$), siguiendo la regla de conteo $(2)(2) = 4$, entonces hay cuatro resultados distintos. Como ya se mostró, estos resultados son $S = \{(H, H), (H, T), (T, H), (T, T)\}$. El número de resultados experimentales de seis monedas es $(2)(2)(2)(2)(2)(2) = 64$.

FIGURA 4.2 DIAGRAMA DE ÁRBOL PARA EL LANZAMIENTO DE DOS MONEDAS



Sin el diagrama de árbol podría pensarse que sólo se pueden tener tres resultados experimentales en dos lanzamientos de una moneda: 0 caras, 1 cara y 2 caras.

Un **diagrama de árbol** es una representación gráfica que permite visualizar un experimento de pasos múltiples. En la figura 4.2 aparece un diagrama de árbol para el experimento del lanzamiento de dos monedas. La secuencia de los pasos en el diagrama va de izquierda a derecha. El paso 1 corresponde al lanzamiento de la primera moneda, el paso 2 al de la segunda moneda. En cada paso, los dos resultados posibles son cruz o cara. Observe que a cada uno de los resultados posibles en el paso 1 pertenecen dos ramas por los dos posibles resultados en el paso 2. Cada uno de los puntos en el extremo derecho del árbol representa un resultado experimental. Cada trayectoria a través del árbol, desde el nodo más a la izquierda hasta uno de los nodos en el extremo derecho del árbol, muestra una secuencia única de resultados.

Ahora una aplicación de la regla de conteo para experimentos de pasos múltiples en el análisis de un proyecto de expansión de la empresa Kentucky Power & Light (KP&L). Kentucky Power & Light ha empezado un proyecto que tiene como objetivo incrementar la capacidad de generación de una de sus plantas en el norte de Kentucky. El proyecto fue dividido en dos etapas o pasos sucesivos: etapa 1 (diseño) y etapa 2 (construcción). A pesar de que cada etapa se planeará y controlará con todo el cuidado posible, a los administrativos no les es posible pronosticar el tiempo exacto requerido en cada una de las etapas del proyecto. En un análisis de proyectos de construcción similares encuentran que la posible duración de la etapa de diseño es de 2, 3, o 4 meses y que la duración de la construcción es de 6, 7 u 8 meses. Además, debido a la necesidad urgente de más energía eléctrica, los administrativos han establecido como meta 10 meses para la terminación de todo el proyecto.

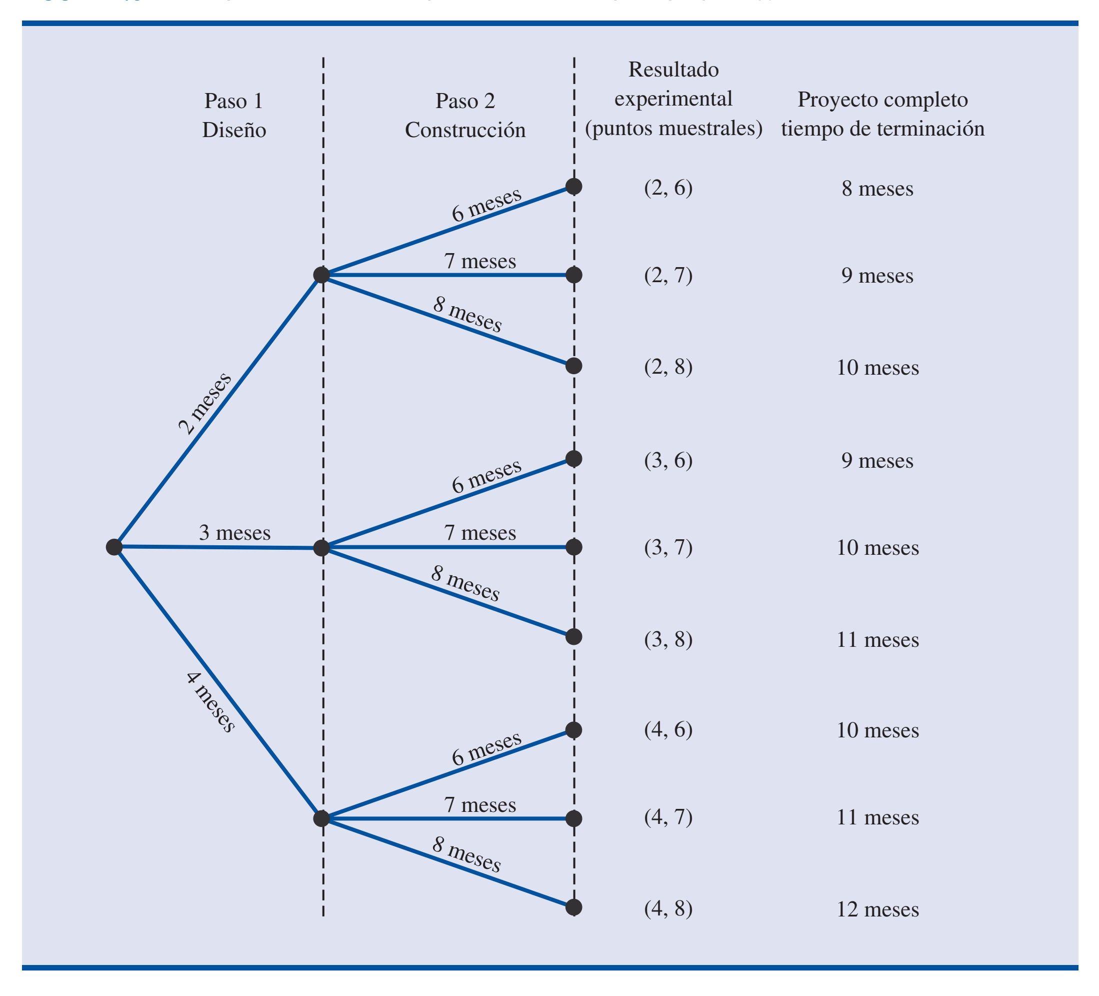
Como hay tres posibles periodos para la etapa del diseño (paso 1) y tres para la etapa de la construcción (paso 2) cabe aplicar la regla de conteo para experimentos de pasos múltiples, entonces el total de resultados posibles es $(3)(3) = 9$. Para describir los resultados experimentales emplean una notación de dos números; por ejemplo, (2, 6) significa que la etapa del diseño durará 2 meses y la etapa de la construcción 6. Esto da como resultado una duración de $2 + 6 = 8$ meses para todo el proyecto. En la tabla 4.1 aparecen los nueve resultados experimentales que hay para el problema de KP&L. El diagrama de árbol de la figura 4.3 muestra como se presentan los nueve resultados (puntos muestrales).

La regla de conteo y el diagrama de árbol ayudan al administrador del proyecto a identificar los resultados experimentales y a determinar la posible duración del proyecto. De acuerdo con la

TABLA 4.1 RESULTADOS EXPERIMENTALES (PUNTOS MUESTRALES) PARA EL PROYECTO KP&L

Duración (meses)			
Etapa 1 Diseño	**Etapa 2 Construcción**	**Notación para los resultados experimentales**	**Proyecto completo: duración (meses)**
2	6	(2, 6)	8
2	7	(2, 7)	9
2	8	(2, 8)	10
3	6	(3, 6)	9
3	7	(3, 7)	10
3	8	(3, 8)	11
4	6	(4, 6)	10
4	7	(4, 7)	11
4	8	(4, 8)	12

FIGURA 4.3 DIAGRAMA DE ÁRBOL PARA EL PROYECTO KP&L



información de la figura 4.3, la duración del proyecto es de 8 a 12 meses, y seis de los nueve resultados experimentales tienen la duración deseada de 10 meses o menos. Aun cuando identificar los resultados experimentales ayuda, es necesario considerar cómo asignar los valores de probabilidad a los resultados experimentales antes de evaluar la probabilidad de que el proyecto dure los 10 meses deseados.

Combinaciones Otra regla de conteo útil le permite contar el número de resultados experimentales cuando el experimento consiste en seleccionar n objetos de un conjunto (usualmente mayor) de N objetos. Ésta es la regla de conteo para combinaciones.

REGLA DE CONTEO PARA COMBINACIONES

El número de combinaciones de N objetos tomados de n en n es

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{4.1}$$

donde

$$N! = N(N-1)(N-2)\cdots(2)(1)$$
$$n! = n(n-1)(n-2)\cdots(2)(1)$$

y por definición,

$$0! = 1$$

La notación ! significa *factorial;* por ejemplo, 5 factorial es $5! = (5)(4)(3)(2)(1) = 120$.

Cuando se hace un muestreo de una población finita de tamaño N, la regla de conteo para combinaciones sirve para hallar el número de muestras de tamaño n que pueden seleccionarse.

Como ejemplo del uso de la regla de conteo para combinaciones, considere un procedimiento de control de calidad en el que un inspector selecciona al azar dos de cinco piezas para probar que no tengan defectos. En un conjunto de cinco partes, ¿cuántas combinaciones de dos partes pueden seleccionarse? De acuerdo con la regla de conteo de la ecuación (4.1) es claro que con $N = 5$ y $n = 2$ se tiene

$$C_2^5 = \binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{(5)(4)(3)(2)(1)}{(2)(1)(3)(2)(1)} = \frac{120}{12} = 10$$

De manera que hay 10 resultados posibles en este experimento de la selección aleatoria de dos partes de un conjunto de cinco. Si etiqueta dichas partes como A, B, C, D y E, las 10 combinaciones o resultados experimentales serán AB, AC, AD, AE, BC, BD, BE, CD, CE y DE.

Para ver otro ejemplo, considere la lotería de Florida en la que se seleccionan seis números de un conjunto de 53 números para determinar al ganador de la semana. Para establecer las distintas variables en la selección de seis enteros de un conjunto de 53, se usa la regla de conteo para combinaciones.

$$\binom{53}{6} = \frac{53!}{6!(53-6)!} = \frac{53!}{6!47!} = \frac{(53)(52)(51)(50)(49)(48)}{(6)(5)(4)(3)(2)(1)} = 22\,957\,480$$

La regla de conteo para combinaciones muestra que la probabilidad de ganar en esta lotería es muy pequeña.

La regla de conteo para combinaciones arroja casi 23 millones de resultados experimentales en esta lotería. Si una persona compra un billete de lotería, tiene una en 22 957 480 posibilidades de ganar la lotería.

Permutaciones La tercera regla de conteo que suele ser útil, es para permutaciones. Dicha regla permite calcular el número de resultados experimentales cuando se seleccionan n objetos de

un conjunto de N objetos y el orden de selección es relevante. Los mismos n objetos seleccionados en orden diferente se consideran un resultado experimental diferente.

REGLA DE CONTEO PARA PERMUTACIONES

El número de permutaciones de N objetos tomados de n en n está dado por

$$P_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \tag{4.2}$$

La regla de conteo para permutaciones tiene relación estrecha con la de combinaciones; sin embargo, con el mismo número de objetos, el número de permutaciones que se obtiene en un experimento es mayor que el número de combinaciones, ya que cada selección de n objetos se ordena de $n!$ maneras diferentes.

Para ver un ejemplo, reconsidere el proceso de control de calidad en el que un inspector selecciona dos de cinco piezas para probar que no tienen defectos. ¿Cuántas permutaciones puede seleccionar? La ecuación (4.2) indica que si $N = 5$ y $n = 2$, se tiene

$$P_2^5 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{(5)(4)(3)(2)(1)}{(3)(2)(1)} = \frac{120}{6} = 20$$

De manera que el experimento de seleccionar aleatoriamente dos piezas de un conjunto de cinco piezas, teniendo en cuenta el orden en que se seleccionen, tiene 20 resultados. Si las piezas se etiquetan A, B, C, D y E, las 20 permutaciones son AB, BA, AC, CA, AD, DA, AE, EA, BC, CB, BD, DB, BE, EB, CD, DC, CE, EC, DE y ED.

Asignación de probabilidades

Ahora verá cómo asignar probabilidades a los resultados experimentales. Los tres métodos comúnmente usados son el método clásico, el método de la frecuencia relativa y el método subjetivo. Sin importar el método que se use, es necesario satisfacer los **requerimientos básicos para la asignación de probabilidades**.

REQUERIMIENTOS BÁSICOS PARA LA ASIGNACIÓN DE PROBABILIDADES

1. La probabilidad asignada a cada resultado experimental debe estar entre 0 y 1, inclusive. Si denota con E_i el i-ésimo resultado experimental y con $P(E_i)$ su probabilidad, entonces exprese este requerimiento como

$$0 \leq P(E_i) \leq 1 \text{ para toda } i \tag{4.3}$$

2. La suma de las probabilidades de los resultados experimentales debe ser igual a 1.0. Para resultados experimentales n escriba este requerimiento como

$$P(E_1) + P(E_2) + \cdots + P(E_n) = 1 \tag{4.4}$$

El **método clásico** de asignación de probabilidades es apropiado cuando todos los resultados experimentales tienen la misma posibilidad. Si existen n resultados experimentales, la probabilidad asignada a cada resultado experimental es $1/n$. Cuando emplee este método, satisfará en automático los dos requerimientos básicos de la asignación de probabilidades.

Por ejemplo, considere el experimento del lanzamiento de una moneda, los dos resultados experimentales —cruz o cara— tienen la misma posibilidad. Como uno de los dos resultados igualmente posibles es cara, la probabilidad de que caiga cara es 1/2 o 0.50. Asimismo, la probabilidad de que caiga cruz también es 1/2 o 0.50.

Otro ejemplo, considere el experimento de lanzar un dado. Es razonable pensar que los seis resultados que pueden presentarse son igualmente posibles y, por tanto, la probabilidad asignada a cada resultado es 1/6. Si $P(1)$ denota la probabilidad de que la cara del dado que caiga hacia arriba sea la que tiene un punto, entonces $P(1) = 1/6$. De manera similar $P(2) = 1/6$, $P(3) = 1/6$, $P(4) = 1/6$, $P(5) = 1/6$ y $P(6) = 1/6$. Observe que dichas probabilidades satisfacen los dos requerimientos básicos de las ecuaciones (4.3) y (4.4), porque cada una es mayor o igual que cero y juntas suman 1.0.

El **método de frecuencia relativa** para la asignación de probabilidades es el más conveniente cuando existen datos para estimar la proporción de veces que se presentarán los resultados si el experimento se repite muchas veces. Considere, por ejemplo un estudio sobre los tiempos de espera en el departamento de rayos *x* de un hospital pequeño. Durante 20 días sucesivos un empleado registra el número de personas que están esperando el servicio a las 9:00 a.m.; los resultados son los siguientes.

Número de personas que esperan	Número de días: resultados de ocurrencia
0	2
1	5
2	6
3	4
4	3
Total	20

En estos datos aparece que 2 de los 20 días, había cero pacientes esperando el servicio, 5 días había un paciente en espera y así sucesivamente. Con el método de la frecuencia relativa, la probabilidad que se le asignará al resultado experimental cero pacientes esperan el servicio, será $2/20 = 0.10$; al resultado experimental un paciente espera el servicio, $5/20 = 0.25$; $6/20 = 0.30$ a dos pacientes esperan el servicio; $4/20 = 0.20$ a tres pacientes esperan el servicio y $3/20 = 0.15$ a cuatro pacientes esperan el servicio. Como sucede con el método clásico, al usar el método de frecuencia relativa se satisfacen en automático los dos requerimientos básicos correspondientes a las ecuaciones (4.3) y (4.4).

El **método subjetivo** de asignación de probabilidades es el más indicado cuando no es factible suponer que todos los resultados de un experimento sean igualmente posibles y, además, cuenta con pocos datos relevantes. El método subjetivo de asignación de probabilidades a los resultados de un experimento, usa toda la información disponible, por ejemplo, la propia experiencia o la intuición. Después de considerar dicha información se asigna un valor de probabilidad que expresa el *grado de confianza* (en una escala de 0 a 1) que tiene acerca de que un resultado experimental ocurra. Como la probabilidad subjetiva expresa el grado de confianza que tiene un individuo, es personal. Cuando se usa el método de probabilidad subjetiva, es de esperarse que personas distintas asignen probabilidades diferentes a los mismos resultados de un experimento.

En el método subjetivo hay que tener cuidado de que se satisfagan los dos requerimientos básicos expresados en las ecuaciones (4.3) y (4.4). Sea cual sea el grado de confianza que tenga la persona, el valor de probabilidad asignado a cada resultado experimental debe estar entre 0 y 1, inclusive, y la suma de las probabilidades de todos los resultados experimentales debe ser 1.0.

Considere el caso en el que Tom y Judy Elsbernd hacen una oferta para la compra de una casa. Hay dos resultados posibles:

$$E_1 = \text{su oferta será aceptada}$$
$$E_2 = \text{su oferta no será aceptada}$$

El teorema de Bayes (véase sección 4.5) proporciona un medio para combinar la probabilidad a priori determinada subjetivamente con probabilidades obtenidas por otros medios para obtener probabilidades a posteriori o revisadas.

Judy cree que la probabilidad de que su oferta sea aceptada es 0.8; por tanto, Judy establece que $P(E_1) = 0.8$ y $P(E_2) = 0.2$; Tom, por su parte, cree que la probabilidad de que su oferta sea aceptada es 0.6; por tanto, Tom establecerá $P(E_1) = 0.6$ y $P(E_2) = 0.4$. Observe que la estimación de probabilidad de E_1 que hace Tom refleja bastante pesimismo de que su oferta sea aceptada.

Tanto Judy como Tom asignaron probabilidades que satisfacen los dos requerimientos básicos. El hecho de que sus probabilidades sean diferentes subraya la naturaleza personal del método subjetivo.

Incluso en situaciones de negocios en que es posible emplear el método clásico o el de las probabilidades relativas, los administradores suelen proporcionar estimaciones subjetivas de una probabilidad. En tales casos, la mejor estimación de una probabilidad suele obtenerse combinando las estimaciones del método clásico o del método de las frecuencias relativas con las estimaciones subjetivas de una probabilidad.

Probabilidades para el proyecto KP&L

Para continuar con el análisis del proyecto KP&L hay que hallar las probabilidades de los nueve resultados experimentales enumerados en la tabla 4.1. De acuerdo con la experiencia, los administrativos concluyen que los resultados experimentales no son todos igualmente posibles. Por tanto, no emplean el método clásico de asignación de probabilidades. Entonces deciden hacer un estudio sobre la duración de los proyectos similares realizados por KP&L en los últimos tres años. En la tabla 4.2 se resume el resultado de este estudio considerando 40 proyectos similares.

Después de analizar los resultados de este estudio, los administrativos deciden emplear el método de frecuencia relativa para asignar las probabilidades. Los administrativos podrían haber aportado probabilidades subjetivas, pero se dieron cuenta de que el proyecto actual era muy similar a los 40 proyectos anteriores. Así, consideraron que el método de frecuencia relativa sería el mejor.

Si emplea la tabla 4.2 para calcular las probabilidades, observará que el resultado (2, 6) —duración de la etapa 1, 2 meses, y duración de la etapa 2, 6 meses— se encuentra seis veces en los 40 proyectos. Con el método de las frecuencias relativas, la probabilidad signada a este resultado es $6/40 = 0.15$. También el resultado (2, 7) se encuentra seis veces en los 40 proyectos $6/40 = 0.15$. Continuando de esta manera, se obtienen, para los puntos muestrales del proyecto de KP&L, las asignaciones de probabilidad que se muestran en la tabla 4.3. Observe que $P(2, 6)$ representa la probabilidad del punto muestral (2, 6), $P(2, 7)$ representa la probabilidad del punto muestral (2, 7) y así sucesivamente.

TABLA 4.2 DURACIÓN DE 40 PROYECTOS DE KP&L

Duración (meses)				
Etapa 1 Diseño	**Etapa 2 Construcción**	**Punto muestral**		**Número de proyectos que tuvieron esta duración**
2	6	(2, 6)		6
2	7	(2, 7)		6
2	8	(2, 8)		2
3	6	(3, 6)		4
3	7	(3, 7)		8
3	8	(3, 8)		2
4	6	(4, 6)		2
4	7	(4, 7)		4
4	8	(4, 8)		6
			Total	40

TABLA 4.3 ASIGNACIÓN DE PROBABILIDADES PARA EL PROYECTO KP&L, EMPLEANDO EL MÉTODO DE LAS FRECUENCIAS RELATIVAS

Punto muestral	Tiempo de terminación del proyecto	Probabilidad del punto muestral
(2, 6)	8 meses	$P(2, 6) = 6/40 =$ 0.15
(2, 7)	9 meses	$P(2, 7) = 6/40 =$ 0.15
(2, 8)	10 meses	$P(2, 8) = 2/40 =$ 0.05
(3, 6)	9 meses	$P(3, 6) = 4/40 =$ 0.10
(3, 7)	10 meses	$P(3, 7) = 8/40 =$ 0.20
(3, 8)	11 meses	$P(3, 8) = 2/40 =$ 0.05
(4, 6)	10 meses	$P(4, 6) = 2/40 =$ 0.05
(4, 7)	11 meses	$P(4, 7) = 4/40 =$ 0.10
(4, 8)	12 meses	$P(4, 8) = 6/40 =$ 0.15
		Total 1.00

NOTAS Y COMENTARIOS

1. En estadística la noción de experimento difiere un poco del concepto de experimento de las ciencias físicas. En las ciencias físicas, los investigadores suelen realizar los experimentos en laboratorios o en ambientes controlados, con objeto de investigar causas y efectos. En los experimentos estadísticos, la probabilidad determina los resultados. Aun cuando un experimento se repita con exactitud, el resultado puede ser completamente diferente. Debido a esta influencia que tiene la probabilidad sobre los resultados, a los experimentos en estadística también se les conoce como *experimentos aleatorios.*

2. Cuando de una población de tamaño N se extrae una muestra aleatoria sin reemplazarla, se emplea la regla de conteo para combinaciones para calcular la cantidad de muestras de tamaño n que pueden seleccionarse.

Ejercicios

Métodos

1. Un experimento consta de tres pasos; para el primer paso hay tres resultados posibles, para el segundo hay dos resultados posibles y para el tercer paso hay cuatro resultados posibles. ¿Cuántos resultados distintos hay para el experimento completo?



2. ¿De cuántas maneras es posible seleccionar tres objetos de un conjunto de seis objetos? Use las letras A, B, C, D, E y F para identificar a los objetos y enumere todas las combinaciones diferentes de tres objetos.
3. ¿Cuántas permutaciones de tres objetos se pueden seleccionar de un grupo de seis objetos? Use las letras A, B, C, D, E y F para identificar a los objetos y enumere cada una de las permutaciones factibles para los objetos B, D y F.
4. Considere el experimento de lanzar una moneda tres veces.
   a. Elabore un diagrama de árbol de este experimento.
   b. Enumere los resultados del experimento.
   c. ¿Cuál es la probabilidad que le corresponde a cada uno de los resultados?
5. Suponga que un experimento tiene cinco resultados igualmente posibles: E_1, E_2, E_3, E_4 y E_5. Asigne probabilidades a los resultados y muestre que satisfacen los requerimientos expresados por las ecuaciones (4.3) y (4.4). ¿Qué método empleó?



6. Un experimento que tiene tres resultados es repetido 50 veces y se ve que E_1 aparece 20 veces, E_2 13 veces y E_3 17 veces. Asigne probabilidades a los resultados. ¿Qué método empleó?

7. La persona que toma las decisiones asigna las probabilidades siguientes a los cuatro resultados de un experimento: $P(E_1) = 0.10$, $P(E_2) = 0.15$, $P(E_3) = 0.40$ y $P(E_4) = 0.20$. ¿Son válidas estas asignaciones de probabilidades? Argumente.

Aplicaciones

8. En una ciudad las solicitudes de cambio de uso de suelo pasan por un proceso de dos pasos: una revisión por la comisión de planeación y la decisión final tomada por el consejo de la ciudad. En el paso 1 la comisión de planeación revisa la solicitud de cambio de uso de suelo y hace una recomendación positiva o negativa respecto al cambio. En el paso 2 el consejo de la ciudad revisa la recomendación hecha por la comisión de planeación y vota para aprobar o desaprobar el cambio de suelo. Suponga que una empresa dedicada a la construcción de complejos departamentales presenta una solicitud de cambio de uso de suelo. Considere el proceso de la solicitud como un experimento. ¿Cuántos puntos muestrales tiene este experimento? Enumérelos. Construya el diagrama de árbol del experimento.



9. El muestreo aleatorio simple usa una muestra de tamaño n tomada de una población de tamaño N para obtener datos para hacer inferencias acerca de las características de la población. Suponga que, de una población de 50 cuentas bancarias, desea tomar una muestra de cuatro cuentas con objeto de tener información acerca de la población. ¿Cuantas muestras diferentes de cuatro cuentas pueden obtener?



10. El capital de riesgo es una fuerte ayuda para los fondos disponibles de las empresas. De acuerdo con Venture Economics (*Investor's Business Daily*, 28 de abril de 2000) de 2374 desembolsos en capital de riesgo, 1434 son de empresas en California, 390 de empresas en Massachussets, 217 de empresas en Nueva York y 112 de empresas en Colorado. Veintidós por ciento de las empresas que reciben fondos se encuentran en las etapas iniciales de desarrollo y 55% en la etapa de expansión. Suponga que desea tomar en forma aleatoria una de estas empresas para saber cómo son usados los fondos de capital de riesgo.
    a. ¿Cuál es la probabilidad de que la empresa que seleccione sea de California?
    b. ¿De que la empresa no sea de ninguno de los estados citados?
    c. ¿De que la empresa elegida no se encuentre en las etapas iniciales de desarrollo?
    d. Si admite que las empresas en las etapas iniciales de desarrollo tuvieran una distribución homogénea en todo el país, ¿cuántas empresas de Massachussets que reciben fondos de capital de riesgo se encuentran en las etapas iniciales de desarrollo?
    e. La cantidad total de fondos invertidos es $32.4 mil millones. Estime la cantidad destinada a Colorado.

11. La National Highway Traffic Safety Administration (NHTSA) realizó una investigación para saber si los conductores de Estados Unidos están usando sus cinturones de seguridad (Associated Press, 25 de agosto de 2003). Los datos muestrales fueron los siguientes.

	Conductores que emplean el cinturón	
Región	**Sí**	**No**
Noreste	148	52
Oeste medio	162	54
Sur	296	74
Oeste	252	48
Total	858	228

    a. ¿Cuál es la probabilidad de que en Estados Unidos un conductor lleve puesto el cinturón?
    b. Un año antes, la probabilidad en Estados Unidos de que un conductor llevara puesto el cinturón era 0.75. El director de NHTSA, doctor Jeffrey Runge esperaba que en 2003 la probabilidad llegara a 0.78. ¿Estará satisfecho con los resultados del estudio del 2003?

c. ¿Cuál es la probabilidad de que se use el cinturón en las distintas regiones del país? ¿En qué región se usa más el cinturón?
d. En la muestra, ¿qué proporción de los conductores provenía de cada región del país? ¿En qué región se seleccionaron más conductores? ¿Qué región viene en segundo lugar?
e. Si admite que en todas las regiones la cantidad de conductores es la misma, ¿ve usted alguna razón para que la probabilidad estimada en el inciso a sea tan alta? Explique.

12. En Estados Unidos hay una lotería que se juega dos veces por semana en 28 estados, en las Islas Vírgenes y en el Distrito de Columbia. Para jugar, debe comprar un billete y seleccionar cinco números del 1 al 55 y un número del 1 al 42. Para determinar al ganador se sacan 5 bolas blancas entre 55 bolas blancas y una bola roja entre 42 bolas rojas. Quien atine a los cinco números de bolas blancas y al número de la bola roja es el ganador. Ocho trabajadores de una empresa tienen el récord del mayor premio, ganaron $365 millones al atinarle a los números 15-17-43-44-49 de las bolas blancas y al 29 de las bolas rojas. En cada juego hay también otros premios. Por ejemplo, quien atina a los cinco números de las bolas blancas se lleva un premio de $200 000 (www.powerball.com, 19 de marzo de 2006).
    a. ¿De cuántas maneras se pueden seleccionar los primeros cinco números?
    b. ¿Cuál es la probabilidad de ganar los $200 000 atinándole a los cinco números de bolas blancas?
    c. ¿Cuál es la probabilidad de atinarle a todos los números y ganar el premio mayor?

13. Una empresa que produce pasta de dientes está analizando el diseño de cinco empaques diferentes. Suponiendo que existe la misma posibilidad de que los clientes elijan cualquiera de los empaques, ¿cuál es la probabilidad de selección que se le asignaría a cada diseño de empaque? En un estudio, se pidió a 100 consumidores que escogieran el diseño que más les gustara. Los resultados se muestran en la tabla siguiente. ¿Confirman estos datos la creencia de que existe la misma posibilidad de que los clientes elijan cualquiera de los empaques? Explique

Diseño	**Número de veces que fue elegido**
1	5
2	15
3	30
4	40
5	10



4.2 Eventos y sus probabilidades

En la introducción de este capítulo el término *evento* fue aplicado tal como se usa en el lenguaje cotidiano. Después, en la sección 4.1 se presentó el concepto de experimento y de los correspondientes resultados experimentales o puntos muestrales. Puntos muestrales y eventos son la base para el estudio de la probabilidad. Por tanto, ahora se le presenta la definición formal de **evento** como se emplea en relación con los puntos muestrales. Con esto se tiene la base para poder dar probabilidades de eventos.

EVENTO

Un evento es una colección de puntos muestrales.

Para dar un ejemplo recuerde el proyecto de KP&L. Considere que al encargado del proyecto le interesa conocer la probabilidad de terminar el proyecto en 10 meses o menos. En la tabla 4.3 aparecen los puntos muestrales (2, 6), (2, 7), (2, 8), (3, 6), (3, 7), (4, 6) correspondientes a una duración del proyecto de 10 meses o menos. C denota el evento de que el proyecto dura 10 meses o menos:

$$C = \{(2, 6), (2, 7), (2, 8), (3, 6), (3, 7), (4, 6)\}$$

Si cualquiera de estos puntos muestrales es el resultado experimental, entonces ocurre el evento C.

Otros eventos de posible interés para el administrador del proyecto KP&L son los siguientes:

L = El evento de que el proyecto esté acabado en *menos* de 10 meses

M = El evento de que el proyecto esté acabado en *más* de 10 meses

De acuerdo con la tabla 4.3 dichos eventos consisten de los siguientes puntos muestrales

$$L = \{(2, 6), (2, 7), (3, 6)\}$$
$$M = \{(3, 8), (4, 7), (4, 8)\}$$

Para el proyecto KP&L existen otros muchos eventos, pero todos serán una colección de puntos muestrales del experimento.

Dadas las probabilidades de los puntos muestrales que se presentan en la tabla 4.3, para calcular la probabilidad de cualquier evento que interese al administrador del proyecto KP&L, se emplea la definición siguiente.

PROBABILIDAD DE UN EVENTO

La probabilidad de cualquier evento es igual a la suma de las probabilidades de los puntos muestrales que forman el evento.

De acuerdo con esta definición, la probabilidad de un determinado evento se calcula sumando las probabilidades de los puntos muestrales (resultados experimentales) que forman el evento. Ahora es posible calcular la probabilidad de que el proyecto dure 10 meses o menos. Como este evento está dado por $C = \{(2, 6), (2, 7), (2, 8), (3, 6), (3, 7), (4, 6)\}$, la probabilidad del evento C denotada por $P(C)$ está dada por

$$P(C) = P(2, 6) + P(2, 7) + P(2, 8) + P(3, 6) + P(3, 7) + P(4, 6)$$

Al consultar las probabilidades de los puntos muestrales de la tabla 4.3, se tiene

$$P(C) = 0.15 + 0.15 + 0.05 + 0.10 + 0.20 + 0.05 = 0.70$$

Así, como el evento de que el proyecto dure menos de 10 meses está dado por $L = \{(2, 6), (2, 7), (3, 6)\}$, la probabilidad de este evento será

$$\begin{aligned} P(L) &= P(2, 6) + P(2, 7) + P(3, 6) \\ &= 0.15 + 0.15 + 0.10 = 0.40 \end{aligned}$$

Por último, el evento de que el proyecto dure más de 10 meses está dado por $M = \{(3, 8), (4, 7), (4, 8)\}$ y por tanto

$$\begin{aligned} P(M) &= P(3, 8) + P(4, 7) + P(4, 8) \\ &= 0.05 + 0.10 + 0.15 = 0.30 \end{aligned}$$

Con estas probabilidades, ahora puede informarle al administrador del proyecto KP&L las probabilidades siguientes: que el proyecto dure 10 meses o menos es 0.70; que dure menos de 10 meses es 0.40 y que dure más de 10 meses es 0.30. Este procedimiento para calcular las probabilidades de los eventos aplica para cualquier evento que interese al administrador del proyecto KP&L.

Siempre que se puedan identificar todos los puntos muestrales de un experimento y asignar a cada uno su probabilidad, es factible calcular la probabilidad de un evento usando la definición. Sin embargo, en muchos experimentos la gran cantidad de puntos muestrales hace en extremo difícil, si no imposible, la determinación de los puntos muestrales, así como la asignación de sus probabilidades correspondientes. En las secciones restantes de este capítulo se presentan algunas relaciones básicas de probabilidad útiles para calcular la probabilidad de un evento, sin necesidad de conocer las probabilidades de todos los puntos muestrales.

NOTAS Y COMENTARIOS

1. El espacio muestral S es un evento. Puesto que contiene todos los resultados experimentales, su probabilidad es 1; es decir $P(S) = 1$.
2. Cuando se usa el método clásico para asignar probabilidades, se parte de que todos los resultados experimentales son igualmente posibles. En tales casos la probabilidad de un evento es calculable contando el número de resultados experimentales que hay en el evento y dividiendo el resultado entre el número total de resultados experimentales.

Ejercicios

Métodos

14. Para un experimento hay cuatro resultados que son igualmente posibles: E_1, E_2, E_3 y E_4.
    a. ¿Cuál es la probabilidad de que ocurra E_2?
    b. ¿De que ocurra cualquiera de dos resultados (por ejemplo, E_1 o E_2)?
    c. ¿De que ocurran tres de estos resultados (E_1 o E_2 o E_4)?



15. Considere el experimento de seleccionar un naipe de una baraja con 52 naipes. Cada naipe es un punto muestral y su probabilidad es 1/52.
    a. Enumere los puntos muestrales del evento si selecciona un as.
    b. Enumere los puntos muestrales del evento si selecciona un trébol.
    c. Enumere los puntos muestrales del evento si selecciona una figura (sota, rey o reina).
    d. Halle la probabilidad correspondiente a cada uno de los eventos de los incisos a, b y c.

16. Considere el experimento que consiste en lanzar un par de dados. Suponga que lo relevante es la suma de los puntos en las dos caras que caen hacia arriba.
    a. ¿Cuántos puntos muestrales habrá? (*Sugerencia:* Use la regla de conteo para experimentos de pasos múltiples.)
    b. Enumere los puntos muestrales.
    c. ¿Cuál es la probabilidad de obtener un 7?
    d. ¿De obtener un 9 o un número mayor?
    e. Como en cada lanzamiento son factibles seis valores pares (2, 4, 6, 8, 10, y 12) y sólo cinco impares (3, 5, 7, 9 y 11), se tendrán más veces resultados pares que impares. ¿Está de acuerdo? Explique
    f. ¿Qué método usó para calcular las probabilidades pedidas?

Aplicaciones



17. Consulte las tablas 4.2 y 4.3 que muestran los puntos muestrales del proyecto KP&L y sus probabilidades.
    a. La etapa del diseño (etapa 1) saldrá del presupuesto si su duración es mayor a 4 meses. Liste los puntos muestrales del evento si la etapa del diseño sale del presupuesto.
    b. ¿Cuál es la probabilidad de que la etapa del diseño salga del presupuesto?
    c. La etapa de la construcción (etapa 2) saldrá del presupuesto si su duración es mayor a 8 meses. Enumere los puntos muestrales del evento si la etapa de construcción sale del presupuesto.
    d. ¿Cuál es la probabilidad de que la etapa de construcción salga del presupuesto?
    e. ¿Cuál es la probabilidad de que las dos etapas salgan del presupuesto?
18. Suponga que el administrador de un complejo grande de departamentos proporciona la siguiente estimación de probabilidades subjetivas acerca del número de departamentos libres que habrá el mes próximo.

Departamentos libres	Probabilidad
0	0.05
1	0.15
2	0.35
3	0.25
4	0.10
5	0.10

    Dé la probabilidad de cada uno de los eventos siguientes.
    a. No haya departamentos libres.
    b. Haya por lo menos 4 departamentos libres.
    c. Haya 2 o menos departamentos libres.
19. Una asociación deportiva realiza un sondeo entre las personas mayores a 6 años respecto de su participación en actividades deportivas. (*Statistical Abstract of the United States: 2002*). El total de la población de estas edades fue 248.5 millones, de los cuales 120.9 millones eran hombres y 127.6 millones mujeres. A continuación se presenta el número de participantes en los cinco deportes principales.

	Participantes (en millones)	
Actividad	**Hombres**	**Mujeres**
Andar en bicicleta	22.2	21.0
Acampar	25.6	24.3
Caminar	28.7	57.7
Hacer ejercicio con aparatos	20.4	24.4
Nadar	26.4	34.4

    a. Estime la probabilidad de que una mujer, elegida al azar, participe en cada una de estas actividades deportivas.
    b. Estime la probabilidad de que un hombre, elegido en forma aleatoria, participe en cada una de estas actividades deportivas.
    c. Estime la probabilidad de que una persona, elegida en forma aleatoria, haga ejercicio caminando.
    d. Suponga que acaba de ver una persona que pasa caminando para hacer ejercicio. ¿Cuál es la probabilidad de que sea mujer?, ¿de que sea hombre?

20. La revista *Fortune* publica anualmente una lista de las 500 empresas más grandes de Estados Unidos. A continuación se presentan los cinco estados en los que hay más de estas 500 empresas de *Fortune*.

Estado	Número de empresas
Nueva York	54
California	52
Texas	48
Illinois	33
Ohio	30

Suponga que se elige una de las 500 empresas de *Fortune*. ¿Cuál es la probabilidad de cada uno de los eventos siguientes?

a. Sea N el evento: la empresa se encuentra en Nueva York. Halle $P(N)$.
b. Sea T el evento: la empresa se encuentra en Texas. Halle $P(T)$.
c. Sea B el evento: la empresa se encuentra en uno de estos cinco estados. Halle $P(B)$.

21. En la tabla siguiente se dan las edades de la población de Estados Unidos (*The World Almanac 2004*). Los datos aparecen en millones de personas.

Edad	Cantidad
19 y menos	80.5
20 a 24	19.0
25 a 34	39.9
35 a 44	45.2
45 a 54	37.7
55 a 64	24.3
65 y más	35.0

Suponga una selección aleatoria de una persona de esta población.

a. ¿Cuál es la probabilidad de que la persona tenga entre 20 y 24 años?
b. ¿De que la persona tenga entre 20 y 34 años?
c. ¿De que tenga 45 años o más?



Algunas relaciones básicas de probabilidad

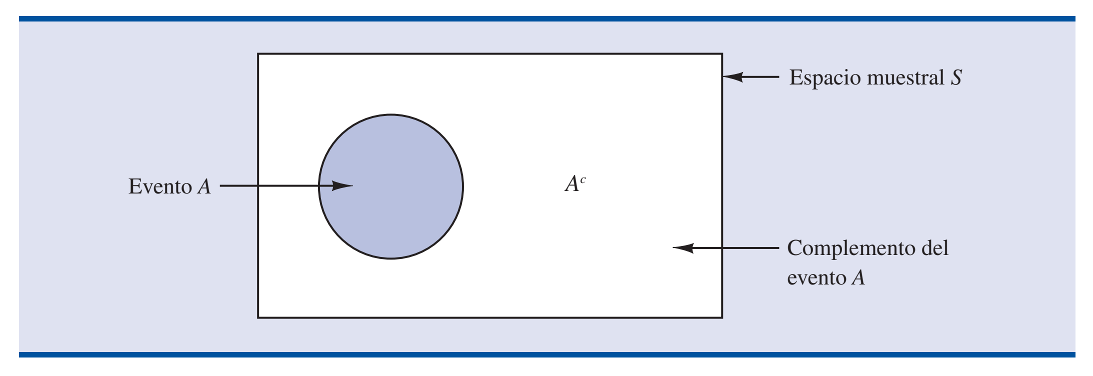
Complemento de un evento

Dado un evento A, el **complemento de A** se define como el evento que consta de todos los puntos muestrales que *no* están en A. El complemento de A se denota A^c. Al diagrama de la figura 4.4 se le llama **diagrama de Venn** e ilustra el concepto del complemento. El área rectangular representa el espacio muestral del experimento y, por tanto, contiene todos los puntos muestrales. El círculo representa el evento A y encierra sólo los puntos muestrales que pertenecen a A. La región del rectángulo que aparece sombreada incluye todos los puntos muestrales que no están en el evento A y es, por definición, el complemento de A.

En cualquier aplicación de la probabilidad ocurre un evento A o su complemento A^c. Por tanto,

$$P(A) + P(A^c) = 1$$

FIGURA 4.4 EL COMPLEMENTO DEL EVENTO A ES EL ÁREA QUE APARECE SOMBREADA



Despejando $P(A)$, obtiene lo siguiente.

CÁLCULO DE UNA PROBABILIDAD USANDO EL COMPLEMENTO

$$P(A) = 1 - P(A^c) \tag{4.5}$$

La ecuación (4.5) indica que la probabilidad de un evento A se puede calcular si se conoce la probabilidad de su complemento, $P(A^c)$.

Por ejemplo, considere el caso de un administrador de ventas que, después de revisar los informes de ventas, encuentra que 80% de los contactos con clientes nuevos no producen ninguna venta. Si A denota el evento hubo venta y A^c el evento no hubo venta, el administrador tiene que $P(A^c) = 0.80$. Mediante la ecuación (4.5) se ve que

$$P(A) = 1 - P(A^c) = 1 - 0.80 = 0.20$$

La conclusión es que la probabilidad de una venta en el contacto con un cliente nuevo es 0.20.

Otro ejemplo, un gerente de compras encuentra que la probabilidad de que el proveedor surta un pedido sin piezas defectuosas es 0.90, empleando el complemento podemos concluir que la probabilidad de que el pedido contenga piezas defectuosas es de $1 - 0.90 = 0.10$.

Ley de la adición

La ley de la adición sirve para determinar la probabilidad de que ocurra por lo menos uno de dos eventos. Es decir, si A y B son eventos, nos interesa hallar la probabilidad de que ocurra el evento A o el B o ambos.

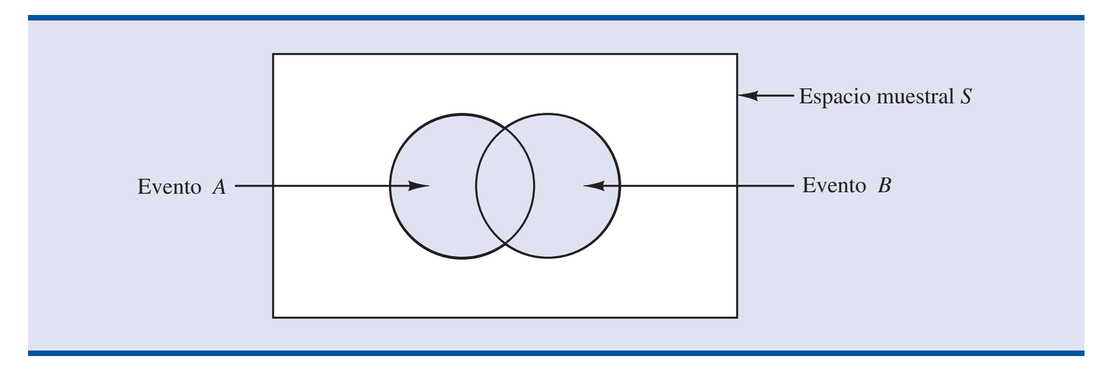
Antes de presentar la ley de la adición es necesario ver dos conceptos relacionados con la combinación de eventos: la *unión* y la *intersección* de eventos. Dados dos eventos, A y B, la **unión de A y B** se define.

UNIÓN DE DOS EVENTOS

La unión de A y B es el evento que contiene todos los puntos muestrales que pertenecen a A o a B o a ambos. La unión se denota $A \cup B$.

El diagrama de Venn de la figura 4.5 representa la unión de los eventos A y B. Observe que en los dos círculos están contenidos todos los puntos muestrales del evento A y todos los puntos

FIGURA 4.5 LA UNIÓN DE LOS EVENTOS *A* Y *B* APARECE SOMBREADA



muestrales del evento *B*. El que los círculos se traslapen indica que algunos puntos muestrales están contenidos tanto en *A* como en *B*.

A continuación la definición de la **intersección de *A* y *B***:

INTERSECCIÓN DE DOS EVENTOS

Dados dos eventos *A* y *B*, la intersección de *A* y *B* es el evento que contiene los puntos muestrales que pertenecen tanto a *A* como a *B*.

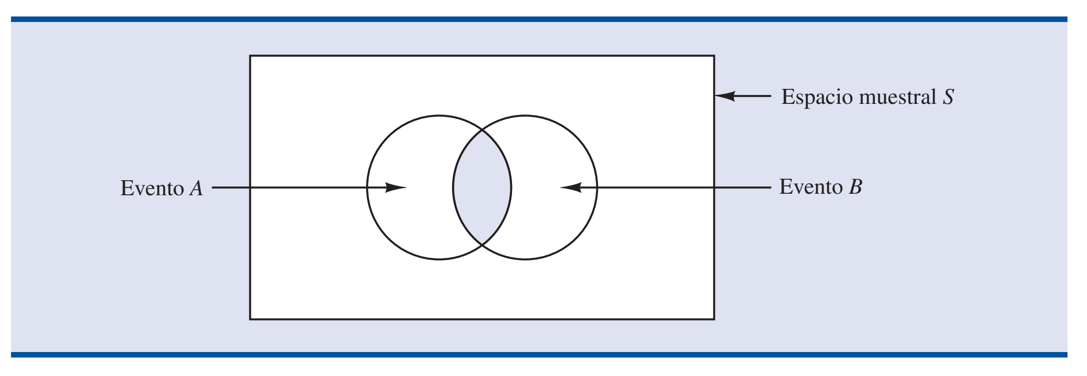
El diagrama de Venn ilustra la intersección de los eventos *A* y *B* mostrados en la figura 4.6. El área donde los círculos se sobreponen es la intersección que contiene una muestra de los puntos que están tanto en *A* como en *B*.

Ahora ya puede continuar con la ley de la adición. La **ley de la adición** proporciona una manera de calcular la probabilidad de que ocurra el evento *A* o el evento *B* o ambos. En otras palabras, la ley de la adición se emplea para calcular la probabilidad de la unión de los dos eventos. La ley de la adición se expresa.

LEY DE LA ADICIÓN

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.6}$$

FIGURA 4.6 LA INTERSECCIÓN DE LOS EVENTOS *A* Y *B* APARECE SOMBREADA



Para que logre un entendimiento intuitivo de la ley de la adición, observe que en la ley de la adición, los dos primeros términos $P(A) + P(B)$, corresponden a los puntos muestrales en $A \cup B$. Pero, como los puntos muestrales que se encuentran en la intersección $A \cap B$ están tanto en A como en B, cuando se calcula $P(A) + P(B)$, los puntos que se encuentran en $A \cap B$ cuentan dos veces. Esto se corrige restando $P(A \cap B)$.

Para ver un ejemplo de la aplicación de la ley de la adición, considere el caso de una pequeña empresa de ensamble en la que hay 50 empleados. Se espera que todos los trabajadores terminen su trabajo a tiempo y que pase la inspección final. A veces, alguno de los empleados no satisface el estándar de desempeño, ya sea porque no termina a tiempo su trabajo o porque no ensambla bien una pieza. Al final del periodo de evaluación del desempeño, el jefe de producción encuentra que 5 de los 50 trabajadores no terminarón su trabajo a tiempo, 6 de los 50 trabajadores ensamblaron mal una pieza y 2 de los 50 trabajadores no terminarón su trabajo a tiempo *y* armaron mal una pieza.

Sea

$$L = \text{el evento no se terminó el trabajo a tiempo}$$
$$D = \text{el evento se armó mal la pieza}$$

La información de las frecuencias relativas lleva a las probabilidades siguientes.

$$P(L) = \frac{5}{50} = 0.10$$
$$P(D) = \frac{6}{50} = 0.12$$
$$P(L \cap D) = \frac{2}{50} = 0.04$$

Después de analizar los datos del desempeño, el jefe de producción decide dar una calificación baja al desempeño de los trabajadores que no terminaron a tiempo su trabajo o que armaron mal alguna pieza; por tanto, el evento de interés es $L \cup D$. ¿Cuál es la probabilidad de que el jefe de producción dé a un trabajador una calificación baja de desempeño?

Observe que esta pregunta sobre probabilidad se refiere a la unión de dos eventos. En concreto, se desea hallar $P(L \cup D)$, usando la ecuación (4.6) se tiene

$$P(L \cup D) = P(L) + P(D) - P(L \cap D)$$

Como conoce las tres probabilidades del lado derecho de esta expresión, se tiene

$$P(L \cup D) = 0.10 + 0.12 - 0.04 = 0.18$$

Estos cálculos indican que la probabilidad de que un empleado elegido al azar obtenga una calificación baja por su desempeño es 0.18

Para ver otro ejemplo de la ley de la adición, considere un estudio reciente efectuado por el director de personal de una empresa importante de software. En el estudio encontró que 30% de los empleados que se van de la empresa antes de dos años, lo hacen por estar insatisfechos con el salario, 20% se van de la empresa por estar descontentos con el trabajo y 12% por estar insatisfechos con las *dos* cosas, el salario y el trabajo. ¿Cuál es la probabilidad de que un empleado

que se vaya de la empresa en menos de dos años lo haga por estar insatisfecho con el salario, con el trabajo o con las dos cosas?

Sea

S = el evento el empleado se va de la empresa por insatisfacción con el salario
W = el evento el empleado se va de la empresa por insatisfacción con el trabajo

Se tiene $P(S) = 0.30$, $P(W) = 0.20$ y $P(S \cap W) = 0.12$. Al aplicar la ecuación (4.6), de la ley de la adición, se tiene

$$P(S \cup W) = P(S) + P(W) - P(S \cap W) = 0.30 + 0.20 - 0.12 = 0.38.$$

Así, la probabilidad de que un empleado se vaya de la empresa por el salario o por el trabajo es 0.38.

Antes de concluir el estudio de la ley de la adición se considerará un caso especial que surge cuando los **eventos son mutuamente excluyentes**.

EVENTOS MUTUAMENTE EXCLUYENTES

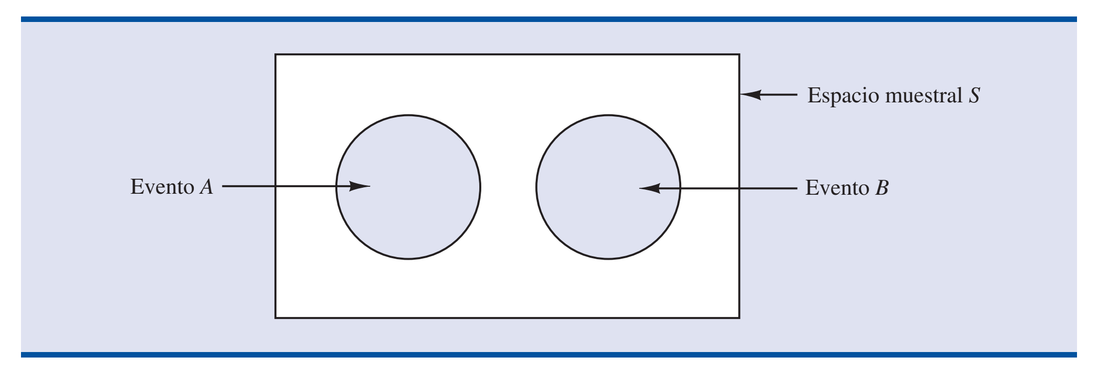
Se dice que dos eventos son mutuamente excluyentes si no tienen puntos muestrales en común.

Los eventos A y B son mutuamente excluyentes si, cuando un evento ocurre, el otro no puede ocurrir. Por tanto, para que A y B sean mutuamente excluyentes, se requiere que su intersección no contenga ningún punto muestral. En la figura 4.7 aparece el diagrama de Venn que representa dos eventos, A y B, mutuamente excluyentes. En este caso $P(A \cap B) = 0$ y la ley de la adición se expresa como sigue:

LEY DE LA ADICIÓN PARA EVENTOS MUTUAMENTE EXCLUYENTES

$$P(A \cup B) = P(A) + P(B)$$

FIGURA 4.7 EVENTOS MUTUAMENTE EXCLUYENTES



Ejercicios

Métodos

22. Suponga que tiene un espacio muestral con cinco resultados experimentales que son igualmente posibles: E_1, E_2, E_3, E_4 y E_5. Sean

$$A = \{E_1, E_2\}$$
$$B = \{E_3, E_4\}$$
$$C = \{E_2, E_3, E_5\}$$

   a. Halle $P(A)$, $P(B)$ y $P(C)$.
   b. Calcule $P(A \cup B)$. ¿A y B son mutuamente excluyentes?
   c. Estime A^c, C^c, $P(A^c)$ y $P(C^c)$.
   d. Halle $A \cup B^c$ y $P(A \cup B^c)$.
   e. Halle $P(B \cup C)$.



23. Suponga que se tiene el espacio muestral $S = \{E_1, E_2, E_3, E_4, E_5, E_6, E_7\}$, donde $E_1, E_2, ..., E_7$ denotan puntos muestrales. La asignación de probabilidades es la siguiente: $P(E_1) = 0.05$, $P(E_2) = 0.20$, $P(E_3) = 0.20$, $P(E_4) = 0.25$, $P(E_5) = 0.15$, $P(E_6) = 0.10$ y $P(E_7) = 0.05$. Sea

$$A = \{E_1, E_4, E_6\}$$
$$B = \{E_2, E_4, E_7\}$$
$$C = \{E_2, E_3, E_5, E_7\}$$

   a. Halle $P(A)$, $P(B)$ y $P(C)$.
   b. Encuentre $A \cup B$ y $P(A \cup B)$.
   c. Halle $A \cap B$ y $P(A \cap B)$.
   d. ¿Los eventos A y B son mutuamente excluyentes?
   e. Halle B^c y $P(B^c)$

Aplicaciones

24. Las autoridades de Clarkson University realizaron un sondeo entre sus alumnos para conocer su opinión acerca de su universidad. Una pregunta fue si la universidad no satisface sus expectativas, si las satisface o si supera sus expectativas. Encontraron que 4% de los interrogados no dieron una respuesta, 26% respondieron que la universidad no llenaba sus expectativas y 56% indicó que la universidad superaba sus expectativas.
   a. Si toma un alumno al azar, ¿cuál es la probabilidad de que diga que la universidad supera sus expectativas?
   b. Si toma un alumno al azar, ¿cuál es la probabilidad de que diga que la universidad satisface o supera sus expectativas?

25. La Oficina de Censos de Estados Unidos cuenta con datos sobre la cantidad de adultos jóvenes, entre 18 y 24 años, que viven en casa de sus padres.* Sea

   M = el evento adulto joven que vive en casa de sus padres
   F = el evento adulta joven que vive en casa de sus padres

   Si toma al azar un adulto joven y una adulta joven, los datos de dicha oficina permiten concluir que $P(M) = 0.56$ y $P(F) = 0.42$ (*The World Almanac,* 2006). La probabilidad de que ambos vivan en casa de sus padres es 0.24.
   a. ¿Cuál es la probabilidad de que al menos uno de dos adultos jóvenes seleccionados viva en casa de sus padres?
   b. ¿Cuál es la probabilidad de que los dos adultos jóvenes seleccionados vivan en casa de sus padres?

*En estos datos se incluye a los adultos jóvenes solteros que viven en los internados de las universidades, porque es de suponer que estos adultos jóvenes vuelven a las casas de sus padres en las vacaciones.

26. Datos sobre las 30 principales acciones y fondos balanceados proporcionan los rendimientos porcentuales anuales y a 5 años para el periodo que termina el 31 de marzo de 2000 (*The Wall Street Journal*, 10 de abril de 2000). Suponga que considera altos un rendimiento anual arriba de 50% y un rendimiento a cinco años arriba de 300%. Nueve de los fondos tienen un rendimiento anual arriba de 50%, siete de los fondos a cinco años lo tienen arriba de 300% y cinco de los fondos tienen tanto un rendimiento anual arriba de 50% como un rendimiento a cinco años arriba de 300%.
   a. ¿Cuál es la probabilidad de un rendimiento anual alto y cuál es la probabilidad de un rendimiento a cinco años alto?
   b. ¿Cuál es la probabilidad de ambos, un rendimiento anual alto y un rendimiento a cinco años alto?
   c. ¿Cuál es la probabilidad de que no haya un rendimiento anual alto ni un rendimiento a cinco años alto?

27. En una encuesta en la pretemporada de futbol americano de la NCAA 2001 se preguntó: "¿Este año habrá un equipo del Big Ten o del Pac-10 en el juego del Rose Bowl?" De los 13 429 interrogados, 2961 dijeron que habría uno del Big Ten, 4494 señalaron que habría uno del Pac-10 y 6823 expresaron que ni el Big Ten ni el Pac-10 tendría un equipo en el Rose Bowl (www.yahoo-.com, 30 de agosto de 2001).
   a. ¿Cuál es la probabilidad de que el interrogado responda que ni el Big Ten ni el Pac-10 tendrán un equipo en el Rose Bowl?
   b. ¿De que afirme que el Big Ten o el Pac-10 tendrán un equipo en el campeonato Rose Bowl?
   c. Halle la probabilidad de que la respuesta sea que tanto el Big Ten como el Pac-10 tendrán un equipo en el Rose Bowl.



28. En una encuesta aplicada a los suscriptores de una revista se encontró que en los últimos 12 meses 45.8% habían rentado un automóvil por razones de trabajo, 54% por razones personales y 30% por razones de trabajo y personales.
   a. ¿Cuál es la probabilidad de que un suscriptor haya rentado un automóvil en los últimos 12 meses por razones de trabajo o por razones personales?
   b. ¿Cuál es la probabilidad de que un suscriptor no haya rentado un automóvil en los últimos 12 meses ni por razones de trabajo ni por razones personales?

29. En Estados Unidos cada año hay más estudiantes con buenas calificaciones que desean inscribirse a las mejores universidades del país. Como el número de lugares permanece relativamente estable, algunas universidades rechazan solicitudes de admisión anticipadas. La universidad de Pensilvania recibió 2851 solicitudes para admisión anticipada. De éstas admitió a 1033 estudiantes, rechazó definitivamente a 854 estudiantes y dejó a 964 para el plazo de admisión normal. Esta universidad admitió a cerca de 18% de los solicitantes en el plazo normal para hacer un total (número de admisiones anticipadas más número de admisiones normales) de 2375 estudiantes (*USA Today* 24 de enero de 2001). Sean los eventos: E, un estudiante que solicita admisión anticipada es admitido; R rechazado definitivamente y D dejado para el plazo normal de admisión, sea A el evento de que un estudiante es admitido en el plazo normal.
   a. Use los datos para estimar $P(E)$, $P(R)$ y $P(D)$.
   b. ¿Son mutuamente excluyentes los eventos E y D? Halle $P(E \cap D)$.
   c. De los 2375 estudiantes admitidos en esta universidad, ¿cuál es la probabilidad de que un estudiante tomado en forma aleatoria haya tenido una admisión anticipada.
   d. Suponga que un estudiante solicita admisión anticipada en esta universidad. ¿Cuál es la probabilidad de que el estudiante tenga una admisión anticipada o en el periodo normal de admisión?

4.4 Probabilidad condicional

Con frecuencia, en la probabilidad de un evento influye el hecho de que un evento relacionado con él ya haya ocurrido. Suponga que tiene un evento A cuya probabilidad es $P(A)$**.** Si obtiene información nueva y sabe que un evento relacionado con él, denotado por B, ya ha ocurrido, de-

seará aprovechar esta información y volver a calcular la probabilidad del evento A. A esta nueva probabilidad del evento A se le conoce como **probabilidad condicional** y se expresa $P(A \mid B)$. La notación | indica que se está considerando la probabilidad del evento A *dada* la condición de que el evento B ha ocurrido. Por tanto, la notación $P(A \mid B)$ se lee "la probabilidad de A dado B".

Como ejemplo de la probabilidad condicional, considere el caso de las promociones de los agentes de policía de una determinada ciudad. La fuerza policiaca consta de 1200 agentes, 960 hombres y 240 mujeres. De éstos, en los últimos dos años, fueron promovidos 340. En la tabla 4.4 se muestra cómo quedaron repartidas estas promociones entre los hombres y mujeres.

Después de analizar el registro de las promociones, un comité femenil protestó, ya que habían sido promovidos 288 agentes hombres, frente a sólo 36 mujeres. Los directivos de la fuerza policiaca argumentaron que el número de mujeres promovidas no se debía a una discriminación, sino a que el número de mujeres que son agentes de policía es una cantidad pequeña. Ahora verá cómo emplear la probabilidad condicional para analizar esta acusación de discriminación.

Sean

M = el evento que un agente de policía sea hombre
W = el evento que un agente de policía sea mujer
A = el evento que un agente de policía sea promovido
A^c = el evento que un agente de policía no sea promovido

Dividir los valores de los datos de la tabla 4.4 entre el total de agentes de policía, 1200, permite concretar la información que se tiene en las probabilidades siguientes.

$P(M \cap A) = 288/1200 = 0.24 =$ probabilidad de que un agente de policía, escogido en forma aleatoria, sea hombre *y* haya sido promovido
$P(M \cap A^c) = 672/1200 = 0.56 =$ probabilidad de que un agente de policía, escogido en forma aleatoria, sea hombre *y* no haya sido promovido
$P(W \cap A) = 36/1200 = 0.03 =$ probabilidad de que un agente de policía, escogido en forma aleatoria, sea mujer *y* haya sido promovido
$P(W \cap A^c) = 204/1200 = 0.17 =$ probabilidad de que un agente de policía, escogido en forma aleatoria, sea mujer *y* no haya sido promovido

Como cada uno de estos valores da la probabilidad de la intersección de dos eventos, se les llama **probabilidades conjuntas**. A la tabla 4.5, que proporciona la información de las probabilidades de promoción de los agentes de policía, se le conoce como *tabla de probabilidades conjuntas*.

Las cantidades que aparecen en los márgenes de una tabla de las probabilidades conjuntas son las probabilidades de cada uno de los eventos por separado. Es decir, $P(M) = 0.80$, $P(W) =$

TABLA 4.4 PROMOCIONES, EN LOS ÚLTIMOS DOS AÑOS, DE LOS AGENTES DE POLICÍA

	Hombre	Mujer	Total
Promovido	288	36	324
No promovido	672	204	876
Total	960	240	1200

TABLA 4.5 TABLA DE PROBABILIDAD CONJUNTA PARA LAS PROMOCIONES

Las probabilidades conjuntas aparecen en el cuerpo de la tabla.

	Hombre (M)	Mujer (W)	Total
Promovido (A)	0.24	0.03	0.27
No promovido (A^c)	0.56	0.17	0.73
Total	0.80	0.20	1.00

Las probabilidades marginales aparecen en los márgenes de la tabla.

0.20, $P(A) = 0.27$, $P(A^c) = 0.73$. A estas probabilidades se les conoce como **probabilidades marginales** por encontrarse en los márgenes de una tabla de probabilidad conjunta.

Observe que las probabilidades marginales se obtienen al sumar las probabilidades conjuntas del renglón o columna correspondiente de la tabla de probabilidades conjuntas. Por ejemplo, la probabilidad marginal de ser promovido es $P(A) = P(M \cap A) + P(W \cap A) = 0.24 + 0.03 = 0.27$. En las probabilidades marginales se observa que 80% de la fuerza policiaca está formada por hombres y 20% por mujeres, que 27% de los agentes de policía fueron promovidos y 73% no fueron promovidos.

Ahora empiece con el análisis de la probabilidad condicional calculando la probabilidad de que un agente de policía sea promovido dado que ese agente sea hombre. Emplee la notación para probabilidad condicional para determinar $P(A \mid M)$. Para calcular $P(A \mid M)$ se observa, primero, que esta notación sólo significa que se considera la probabilidad del evento A (promoción) ya que la condición designada como evento M (que el agente de policía sea hombre) está dada. Así que $P(A \mid M)$ indica que sólo interesan los promovidos dentro de los 960 agentes de policía que son hombres. Como 288 de los 960 agentes de policía que son hombres fueron promovidos, la probabilidad de ser promovido dado que se es un agente hombre es 288/960 = 0.30. En otras palabras, puesto que un agente de policía es hombre, ese agente tuvo 30% de probabilidades de ser promovido en los dos últimos años.

Resultó fácil aplicar este procedimiento, ya que en la tabla 4.4 se muestra el número de agentes de policía en cada categoría. Ahora es interesante mostrar cómo calcular probabilidades condicionales, como $P(A \mid M)$, a partir de las probabilidades de eventos relacionados y no a partir de los datos de frecuencias de la tabla 4.4.

Entonces, $P(A \mid M) = 288/960 = 0.30$. Ahora, tanto el numerador como el denominador de esta fracción se dividen entre 1200, cantidad total de agentes de policía en el estudio.

$$P(A \mid M) = \frac{288}{960} = \frac{288/1200}{960/1200} = \frac{0.24}{0.80} = 0.30$$

Observe que la probabilidad condicional se obtiene de 0.24/0.80. Regrese a la tabla de probabilidad conjunta (tabla 4.5) y observe que 0.24 es la probabilidad conjunta de A y M; es decir, $P(A \cap M) = 0.24$; también que 0.80 es la probabilidad marginal de que un agente de la policía seleccionado aleatoriamente sea hombre. Es decir, $P(M) = 0.80$. Por tanto, la probabilidad condicional $P(A \mid M)$ se calcula como la razón entre $P(A \cap M)$ y la probabilidad marginal $P(M)$.

$$P(A \mid M) = \frac{P(A \cap M)}{P(M)} = \frac{0.24}{0.80} = 0.30$$

El hecho de que la probabilidad condicional se pueda calcular como la razón entre una probabilidad conjunta respecto a una probabilidad marginal proporciona la siguiente fórmula para el cálculo de la probabilidad condicional de dos eventos A y B.

PROBABILIDAD CONDICIONAL

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{4.7}$$

o

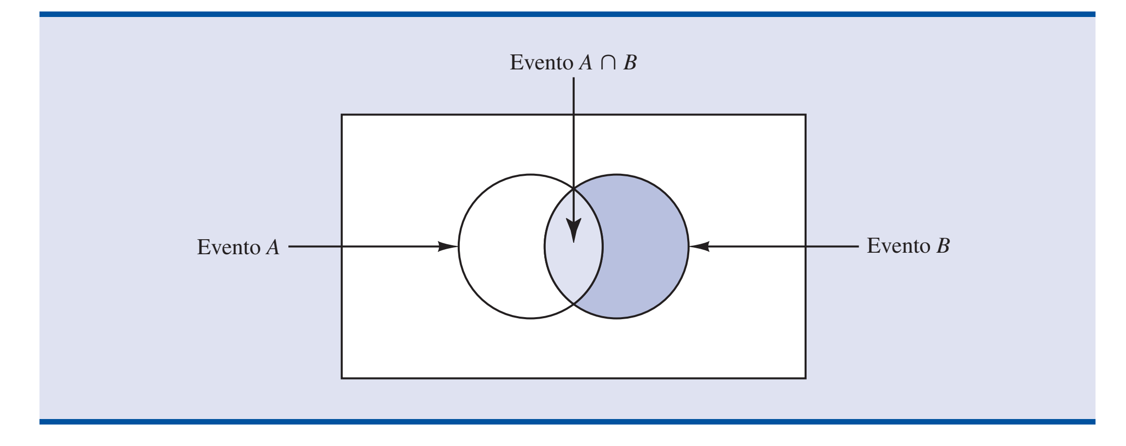
$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{4.8}$$

El diagrama de Venn de la figura 4.8 ayuda a lograr una comprensión intuitiva de la probabilidad condicional. El círculo de la derecha muestra que el evento B ha ocurrido, la parte del círculo que se superpone con el evento A se denota $(A \cap B)$. Una vez que el evento B ha ocurrido, la única manera de que también sea observable el evento A es que ocurra el evento $(A \cap B)$. De manera que la razón $P(A \cap B)/P(B)$ aporta la probabilidad condicional de que se observe el evento A dado que el evento B ya ha ocurrido.

Ahora, considere de nuevo el asunto de la discriminación contra las mujeres agentes de policía. La probabilidad marginal del renglón 1 de la tabla 4.5 indica que la probabilidad de que un agente de la policía sea promovido (ya sea hombre o mujer) es $P(A) = 0.27$. Sin embargo, la cuestión relevante en el caso de la discriminación tiene que ver con las probabilidades condicionales $P(A \mid M)$ y $P(A \mid W)$. Es decir, ¿cuál es la probabilidad de que un agente de la policía sea promovido *dado que* es hombre y cuál es la probabilidad que un agente de la policía sea promovido *dado que* es mujer? Si estas dos probabilidades son iguales, no hay fundamentos para un argumento de discriminación ya que las oportunidades de ser promovidos son las mismas para agentes de la policía hombres o mujeres. Pero, si hay diferencia entre estas dos probabilidades condicionales se confirmará que los hombres y mujeres agentes de policía son considerados de manera distinta cuando se trata de las decisiones para promoverlos.

Ya se determinó que $P(A \mid M) = 0.30$. Ahora use los valores de probabilidad de la tabla 4.5 y la ecuación (4.7) de probabilidad condicional para calcular la probabilidad de que un agente de

FIGURA 4.8 PROBABILIDAD CONDICIONAL $P(A \mid B) = P(A \cap B)|P(B)$



la policía sea promovido dado que es mujer; es decir, $P(A \mid W)$. Use la ecuación (4.7) con W en lugar de B

$$P(A \mid W) = \frac{P(A \cap W)}{P(W)} = \frac{0.03}{0.20} = 0.15$$

¿Qué conclusión obtiene? La probabilidad de que un agente de policía sea promovido dado que es hombre es 0.30, el doble de 0.15, que es la probabilidad de que un agente de policía sea promovido dado que es mujer. Aunque el uso de la probabilidad condicional no demuestra por sí misma que haya discriminación en este caso, los valores de probabilidad condicional confirman el argumento presentado por las mujeres agentes de policía.

Eventos independientes

En el ejemplo anterior, $P(A) = 0.27$, $P(A \mid M) = 0.30$ y $P(A \mid W) = 0.15$. Es claro que a la probabilidad de ser promovido (evento A) le afecta o le influye el que el oficial sea un hombre o una mujer. En concreto, como $P(A \mid M) \neq P(A)$ los eventos A y M son eventos dependientes. Es decir, a la probabilidad del evento A (ser promovido) la altera o le afecta saber que se da el evento M (que el agente sea hombre). De manera similar, como $P(A \mid W) \neq P(A)$, los eventos A y W son *eventos dependientes.* Pero, si la probabilidad de un evento A no cambia por la existencia del evento M *—es decir, si* $P(A \mid M) = P(A)$—, entonces los eventos A y M son **eventos independientes**. Esto lleva a la definición de la independencia de dos eventos.

EVENTOS INDEPENDIENTES

Dos eventos A y B son independientes si

$$P(A \mid B) = P(A) \tag{4.9}$$

o

$$P(B \mid A) = P(B) \tag{4.10}$$

Si no es así, los eventos son dependientes.

Ley de la multiplicación

Mientras que la ley de las suma de probabilidades sirve para calcular la probabilidad de la unión de dos eventos, la ley de la multiplicación es útil para calcular la probabilidad de la intersección de dos eventos. La ley de la multiplicación se basa en la definición de probabilidad condicional. Al despejar en las ecuaciones (4.7) y (4.8) $P(A \cap B)$, se obtiene la **ley de la multiplicación**.

LEY DE LA MULTIPLICACIÓN

$$P(A \cap B) = P(B)P(A \mid B) \tag{4.11}$$

o

$$P(A \cap B) = P(A)P(B \mid A) \tag{4.12}$$

Para ilustrar el uso de la ley de la multiplicación, considere el caso del departamento de circulación de un periódico al que 84% de los hogares de cierta región están suscritos a la edición diaria del periódico. Si D denota el evento un hogar suscrito a la edición diaria, $P(D) = 0.84$. Además, sabe que la probabilidad de que un hogar ya suscrito a la edición diaria se suscriba también a la edición dominical (evento S) es 0.75; esto es, $P(S \mid D) = 0.75$.

¿Cuál es la probabilidad de que un hogar se subscriba a ambas, a la edición diaria y a la dominical? Emplee la ley de la multiplicación y calcule la probabilidad deseada, $P(S \cap D)$.

$$P(S \cap D) = P(D)P(S \mid D) = 0.84(0.75) = 0.63$$

Así, sabe que 63% de los hogares se suscriben a ambas ediciones, a la diaria y a la dominical.

Antes de terminar esta sección hay que considerar el caso especial de la ley de la multiplicación cuando los eventos involucrados son independientes. Recuerde que los eventos A y B son independientes si $P(A \mid B) = P(A)$ o $P(B \mid A) = P(B)$. Por tanto, con las ecuaciones (4.11) y (4.12) obtiene, para el caso especial de eventos independientes, la siguiente ley de la multiplicación.

LEY DE LA MULTIPLICACIÓN PARA EVENTOS INDEPENDIENTES

$$P(A \cap B) = P(A)P(B) \qquad \textbf{(4.13)}$$

Para calcular la probabilidad de la intersección de dos eventos independientes, simplemente se multiplican las probabilidades correspondientes. Observe que la ley de la multiplicación para eventos independientes proporciona otra manera de determinar si dos eventos son independientes. Es decir, si $P(A \cap B) = P(A)P(B)$, entonces A y B son independientes; si $P(A \cap B) \neq P(A)P(B)$, entonces A y B son dependientes.

Como una aplicación de la ley de la multiplicación para eventos independientes considere el caso del jefe de una gasolinería que por experiencia sabe que 80% de los clientes usan tarjeta de crédito al pagar la gasolina. ¿Cuál es la probabilidad de que los dos siguientes clientes paguen la gasolina con tarjeta de crédito? Sean

A = el evento el primer cliente paga con tarjeta de crédito
B = el evento el segundo cliente paga con tarjeta de crédito

entonces el evento que interesa es $A \cap B$. Si no hay ninguna otra información, será razonable suponer que A y B son eventos independientes. Por tanto,

$$P(A \cap B) = P(A)P(B) = (0.80)(0.80) = 0.64$$

Para concluir esta sección, observe que el interés por la probabilidad condicional surgió porque los eventos suelen estar relacionados. En esos casos, los eventos son dependientes y para calcular la probabilidad de estos eventos se usan las fórmulas para probabilidad condicional de las ecuaciones (4.7) y (4.8). Si dos eventos no están relacionados, son independientes; en este caso a las probabilidades de ninguno de los eventos les afecta el hecho de que el otro evento ocurra o no.

NOTAS Y COMENTARIOS

No hay que confundir la noción de eventos mutuamente excluyentes con la de eventos independientes. Dos eventos cuyas probabilidades no son cero, no pueden ser mutuamente excluyentes e independientes. Si uno de los eventos mutuamente excluyentes ocurre, el otro evento no puede ocurrir; por tanto, la probabilidad de que ocurra el otro evento se reduce a cero.

Ejercicios

Métodos



30. Suponga dos eventos, A y B, y que $P(A) = 0.50$, $P(B) = 0.60$ y $P(A \cap B) = 0.40$.
    a. Halle $P(A \mid B)$.
    b. Halle $P(B \mid A)$.
    c. ¿A y B son independientes? ¿Por qué sí o por qué no?

31. Suponga dos eventos, A y B, que son mutuamente excluyentes. Admita, además, que $P(A) = 0.30$ y $P(B) = 0.40$.
    a. Obtenga $P(A \cap B)$.
    b. Calcule $P(A \mid B)$.
    c. Un estudiante de estadística argumenta que los conceptos de eventos mutuamente excluyentes y eventos independientes son en realidad lo mismo y que si los eventos son mutuamente excluyentes deben ser también independientes. ¿Está usted de acuerdo? Use la información sobre las probabilidades para justificar su respuesta.
    d. Dados los resultados obtenidos, ¿qué conclusión sacaría usted acerca de los eventos mutuamente excluyentes e independientes?

Aplicaciones

32. Debido al aumento de los costos de los seguros, en Estados Unidos 43 millones de personas no cuentan con un seguro médico (*Time*, 1 de diciembre de 2003). En la tabla siguiente se muestran datos muestrales representativos de la cantidad de personas que cuentan con seguro médico.

		Seguro médico	
		Sí	**No**
Edad	**18 a 34**	750	170
	35 o mayor	950	130

    a. Con estos datos elabore una tabla de probabilidad conjunta y úsela para responder las preguntas restantes.
    b. ¿Qué indican las probabilidades marginales acerca de la edad de la población de Estados Unidos?
    c. ¿Cuál es la probabilidad de que una persona tomada en forma aleatoria no tenga seguro médico?
    d. Si la persona tiene entre 18 y 34 años, ¿cuál es la probabilidad de que no tenga seguro médico?
    e. Si la persona tiene 34 años o más ¿cuál es la probabilidad de que no tenga seguro médico?
    f. Si la persona no tiene seguro médico, ¿cuál es la probabilidad de que tenga entre 18 y 34 años?
    g. ¿Qué indica esta información acerca del seguro médico en Estados Unidos?

Auto examen

33. Una muestra de estudiantes de la maestría en administración de negocios, arrojó la siguiente información sobre la principal razón que tuvieron los estudiantes para elegir la escuela en donde hacen sus estudios.

		Razones de su elección			
		Calidad de la escuela	**Costo de la escuela**	**Otras**	**Totales**
Tipo de estudiante	**Tiempo completo**	421	393	76	890
	Medio tiempo	400	593	46	1039
	Totales	821	986	122	1929

    a. Con estos datos elabore una tabla de probabilidad conjunta.
    b. Use las probabilidades marginales: calidad de la escuela, costo de la escuela y otras para comentar cuál es la principal razón por la que eligen una escuela.

c. Si es un estudiante de tiempo completo, ¿cuál es la probabilidad de que la principal razón para su elección de la escuela haya sido la calidad de la escuela?
d. Si es un estudiante de medio tiempo, ¿cuál es la probabilidad de que la principal razón para su elección de la escuela haya sido la calidad de la escuela?
e. Si *A* denota el evento es estudiante de tiempo completo y B denota el evento la calidad de la escuela fue la primera razón para su elección, ¿son independientes los eventos *A* y *B*? Justifique su respuesta.

34. La tabla siguiente muestra las probabilidades de los distintos tipos sanguíneo en la población.

	A	**B**	**AB**	**O**
Rh+	0.34	0.09	0.04	0.38
Rh−	0.06	0.02	0.01	0.06

a. ¿Cuál es la probabilidad de que una persona tenga sangre tipo O?
b. ¿De que tenga sangre Rh−?
c. ¿Cuál es la probabilidad de que una persona sea Rh− dado que la persona tiene sangre tipo O?
d. ¿Cuál es la probabilidad de que una persona tenga sangre tipo B dado que es Rh+?
e. ¿Cuál es la probabilidad de que en un matrimonio, los dos sean Rh−?
f. ¿Cuál es la probabilidad de que en un matrimonio, los dos tengan sangre AB?

35. El Departamento de Estadística Laboral de Estados Unidos reúne datos sobre las ocupaciones de las personas entre 25 y 64 años. La tabla siguiente presenta el número de hombres y mujeres (en millones) en cada una de las categorías ocupacionales.

Ocupación	**Hombres**	**Mujeres**
Directivo/Profesional	19 079	19 021
Enseñanza/Ventas/ Administrativo	11 079	19 315
Servicio	4 977	7 947
Producción con precisión	11 682	1 138
Operadores/Obrero	10 576	3 482
Agricultura/Ganadería/Silvicultura/Pesca	1 838	514

a. Desarrolle una tabla de probabilidad conjunta.
b. ¿Cuál es la probabilidad de que un trabajador mujer sea directivo o profesional?
c. ¿Cuál es la probabilidad de que un trabajador hombre esté en producción con precisión?
d. ¿Es la ocupación independiente del género? Justifique su respuesta con el cálculo de la probabilidad.

36. Reggie Miller de los Indiana Pacers tiene el record de la National Basketball Association de más canastas de 3 puntos anotadas en toda una carrera, acertando en 85% de sus tiros (*USA Today,* 22 de enero de 2004). Suponga que ya casi al final de un juego cometen una falta contra él y le conceden dos tiros.
a. ¿Cuál es la probabilidad de que acierte en los dos tiros?
b. ¿De que acierte en por lo menos uno de los dos tiros?
c. ¿De que no acierte en ninguno de los dos tiros?
d. Al final de un juego de básquetbol suele ocurrir que cometan faltas contra un jugador del equipo opuesto para detener el reloj del juego. La estrategia usual es cometer una falta contra el peor tirador del otro equipo. Suponga que el centro de los Indiana Pacers acierta 58% de sus tiros. Calcule para él las probabilidades calculadas en los incisos a, b y c y muestre que hacer una falta intencional contra el centro de los Indiana Pacers es mejor que hacerlo contra Reggie Miller.

37. Visa Card de Estados Unidos estudia con qué frecuencia usan sus tarjetas (de débito y de crédito) los consumidores jóvenes, entre 18 y 24 años. Los resultados del estudio proporcionan las probabilidades siguientes.

- La probabilidad de que un consumidor use su tarjeta al hacer una compra es 0.37.
- Dado que un consumidor usa su tarjeta, la probabilidad de que tenga entre 18 y 24 años es 0.19.
- Puesto que un consumidor usa su tarjeta, la probabilidad de que sea mayor de 24 años es 0.81.

Datos de la Oficina de Censos de Estados Unidos indican que 14% de los consumidores tienen entre 18 y 24 años.

a. Ya que un consumidor tiene entre 18 y 24 años, ¿cuál es la probabilidad de que use su tarjeta?
b. Dado que un consumidor tiene más de 24 años, ¿cuál es la probabilidad de que use su tarjeta?
c. ¿Qué interpretación se le da a las probabilidades de los incisos a y b?
d. ¿Empresas como Visa, Master Card y Discover deben proporcionar tarjetas a los consumidores entre 18 y 24 años, antes de que tengan una historia crediticia? Si no, explique. Si sí, ¿qué restricciones deben poner las empresas a estos consumidores?

38. En un estudio de Morgan Stanley Consumer Research se muestrearon hombres y mujeres y se les preguntó qué preferían tomar: agua de botella o una bebida deportiva como Gatorade o Propel Fitness (*The Atlanta Journal-Constitution*, 28 de diciembre de 2005). Suponga que en el estudio hayan participado 200 hombres y 200 mujeres y que de todos 280 hayan preferido el agua de botella. En el grupo de los que preferían bebidas deportivas, 80 eran hombres y 40 eran mujeres.

Sea

M = el evento el consumidor es hombre
W = el evento el consumidor es mujer
B = el evento el consumidor prefiere agua de botella
S = el evento el consumidor prefiere una bebida deportiva

a. ¿Cuál es la probabilidad de que en este estudio una persona prefiera agua de botella?
b. ¿De que en este estudio una persona prefiera una bebida deportiva?
c. ¿Cuáles son las probabilidades condicionales $P(M \mid S)$ y $P(W \mid S)$?
d. ¿Cuáles son las probabilidades conjuntas $P(M \cap S)$ y $P(W \cap S)$?
e. Dado que un consumidor es hombre, ¿cuál es la probabilidad de que prefiera una bebida deportiva?
f. Ya que un consumidor es mujer, ¿cuál es la probabilidad de que prefiera una bebida deportiva?
g. ¿Depende la preferencia por una bebida deportiva de que el consumidor sea hombre o mujer? Explique usando la información sobre las probabilidades.



Teorema de Bayes

En el estudio de la probabilidad condicional vio que revisar las probabilidades cuando se obtiene más información es parte importante del análisis de probabilidades. Por lo general, se suele iniciar el análisis con una estimación de probabilidad inicial o **probabilidad previa** de los eventos que interesan. Después, de fuentes como una muestra, una información especial o una prueba del producto, se obtiene más información sobre estos eventos. Dada esta nueva información, se modifican o revisan los valores de probabilidad mediante el cálculo de probabilidades revisadas a las que se les conoce como **probabilidades posteriores**. El **teorema de Bayes** es un medio para calcular estas probabilidades. En la figura 4.9 se presentan los pasos de este proceso de revisión de la probabilidad.

FIGURA 4.9 REVISIÓN DE LA PROBABILIDAD USANDO EL TEOREMA DE BAYES

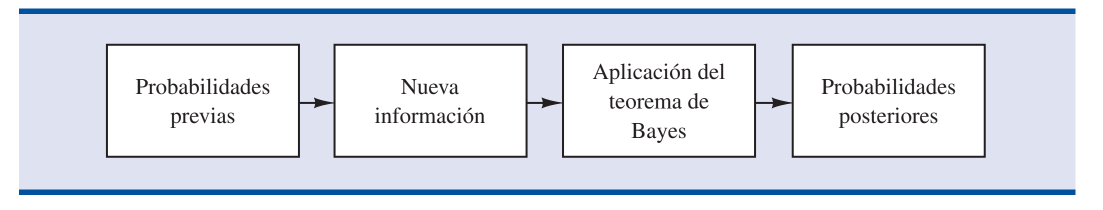


TABLA 4.6 CALIDAD DE DOS PROVEEDORES

	Porcentaje de piezas buenas	Porcentaje de piezas malas
Proveedor 1	98	2
Proveedor 2	95	5

Como aplicación del teorema de Bayes, considere una fábrica que compra piezas de dos proveedores. Sea A_1 el evento la pieza proviene del proveedor 1 y A_2 el evento la pieza proviene del proveedor 2. De las piezas que compra la fábrica, 65% proviene del proveedor 1 y 35% restante proviene del proveedor 2. Por tanto, si toma una pieza aleatoriamente, le asignará las probabilidades previas $P(A_1) = 0.65$ y $P(A_2) = 0.35$.

La calidad de las piezas compradas varía de acuerdo con el proveedor. Por experiencia, sabe que la calidad de los dos proveedores es como muestra la tabla 4.6. Si G denota el evento la pieza está buena y B denota el evento la pieza está mala, la información de la tabla 4.6 proporciona los siguientes valores de probabilidad condicional.

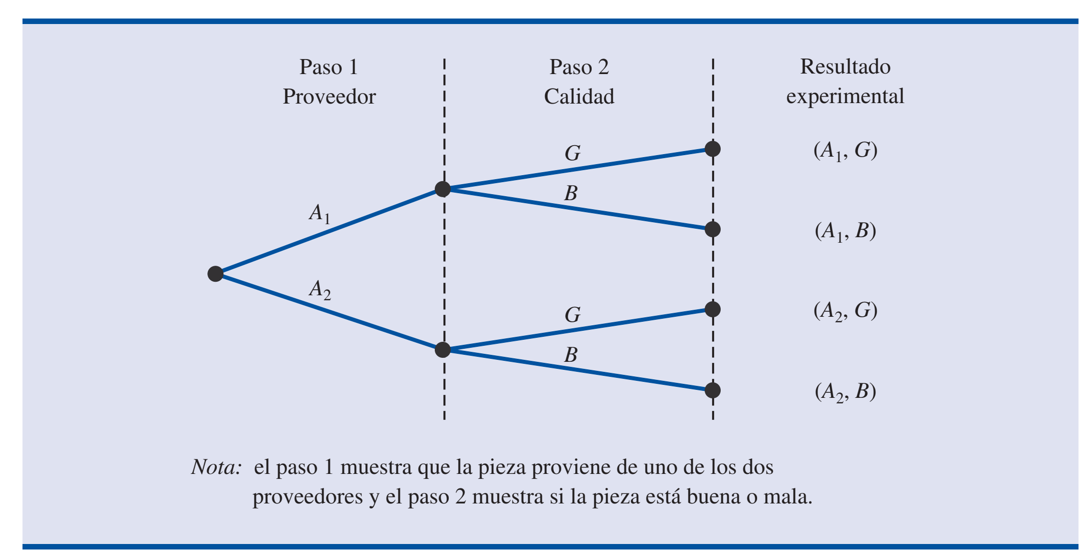
$$P(G \mid A_1) = 0.98 \qquad P(B \mid A_1) = 0.02$$
$$P(G \mid A_2) = 0.95 \qquad P(B \mid A_2) = 0.05$$

El diagrama de árbol de la figura 4.10 representa el proceso de recibir una pieza, de uno de los dos proveedores, y después determinar si la pieza es buena o mala como experimento de dos pasos. Se observa que existen cuatro resultados experimentales: dos corresponden a que la pieza esté buena y dos corresponden a que la pieza esté mala.

Cada uno de los resultados experimentales es la intersección de dos eventos, de manera que para calcular estas probabilidades puede usar la ley de la multiplicación. Por ejemplo,

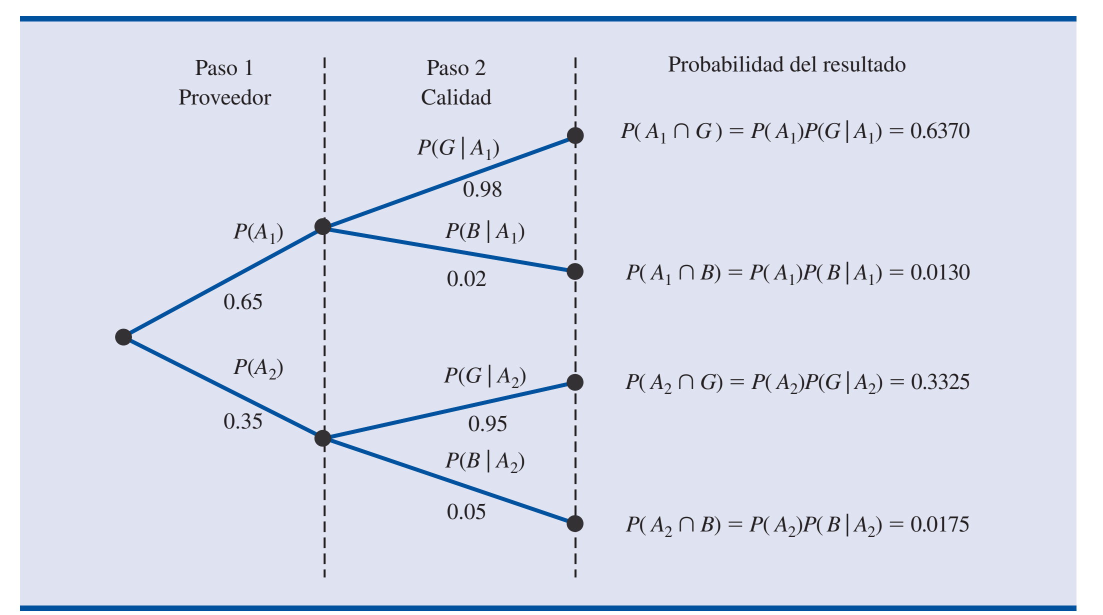
$$P(A_1, G) = P(A_1 \cap G) = P(A_1)P(G \mid A_1)$$

FIGURA 4.10 DIAGRAMA DE ÁRBOL PARA EL EJEMPLO DE LOS DOS PROVEEDORES



Nota: el paso 1 muestra que la pieza proviene de uno de los dos proveedores y el paso 2 muestra si la pieza está buena o mala.

FIGURA 4.11 ÁRBOL DE PROBABILIDAD PARA EL EJEMPLO DE LOS DOS PROVEEDORES



El proceso del cálculo de estas probabilidades conjuntas se representa mediante un árbol de probabilidad (figura 4.11). De izquierda a derecha por el árbol, las probabilidades de cada una de las ramas del paso 1 son probabilidades previas y las probabilidades de cada una de las ramas del paso 2 son probabilidades condicionales. Para hallar la probabilidad de cada uno de los resultados experimentales, simplemente se multiplican las probabilidades de las ramas que llevan a ese resultado. En la figura 4.11 se muestra cada una de estas probabilidades conjuntas junto con las probabilidades en cada rama.

Suponga ahora que las piezas de los dos proveedores se emplean en el proceso de fabricación de esta empresa y que una máquina se descompone al tratar de procesar una pieza mala. Dada la información de que la pieza está mala, ¿cuál es la probabilidad de que sea del proveedor 1 y cuál es la probabilidad de que sea del proveedor 2? Para responder estas preguntas aplique el teorema de Bayes usando la información del árbol de probabilidad (figura 4.11).

Como B es el evento la parte está mala, lo que busca son las probabilidades posteriores $P(A_1 \mid B)$ y $P(A_2 \mid B)$. De acuerdo con la ley para la probabilidad condicional

$$P(A_1 \mid B) = \frac{P(A_1 \cap B)}{P(B)} \qquad \textbf{(4.14)}$$

Del árbol de probabilidad

$$P(A_1 \cap B) = P(A_1)P(B \mid A_1) \qquad \textbf{(4.15)}$$

Para hallar $P(B)$, se observa que B sólo puede presentarse de dos maneras: $(A_1 \cap B)$ y $(A_2 \cap B)$. Por tanto,

$$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) \\ &= P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) \end{aligned} \qquad \textbf{(4.16)}$$

Sustituyendo las ecuaciones (4.15) y (4.16) en la ecuación (4.14) y expresando de manera similar $P(A_2 \mid B)$ se obtiene el teorema de Bayes para el caso de dos eventos.

Al reverendo Thomas Bayes, un ministro presbiteriano, se le atribuye la idea inicial que llevó a la versión del teorema de Bayes que se usa en la actualidad.

TEOREMA DE BAYES (CASO DE DOS EVENTOS)

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \tag{4.17}$$

$$P(A_2 \mid B) = \frac{P(A_2)P(B \mid A_2)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \tag{4.18}$$

A partir de la ecuación (4.17) y los valores de probabilidad del ejemplo, se tiene

$$\begin{aligned} P(A_1 \mid B) &= \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \\ &= \frac{(0.65)(0.02)}{(0.65)(0.02) + (0.35)(0.05)} = \frac{0.0130}{0.0130 + 0.0175} \\ &= \frac{0.0130}{0.0305} = 0.4262 \end{aligned}$$

Y usando la ecuación (4.18) se encuentra $P(A_2 \mid B)$.

$$\begin{aligned} P(A_2 \mid B) &= \frac{(0.35)(0.05)}{(0.65)(0.02) + (0.35)(0.05)} \\ &= \frac{0.0175}{0.0130 + 0.0175} = \frac{0.0175}{0.0305} = 0.5738 \end{aligned}$$

Observe que al principio de este ejemplo, la probabilidad de seleccionar una pieza y que fuera del proveedor 1 era 0.65. Sin embargo, dada la información de que la pieza está mala, la probabilidad de que la pieza provenga del proveedor 1 bajó a 0.4262. En efecto, si la pieza está mala, la posibilidad de que sea del proveedor 2 es mayor que 50-50; es decir, $P(A_2 \mid B) = 0.5738$.

El teorema de Bayes es aplicable cuando los eventos para los que se quiere calcular la probabilidad revisada son mutuamente excluyentes y su unión es todo el espacio muestral.* En el caso de n eventos mutuamente excluyentes $A_1, A_2, \ldots, A_n$, cuya unión sea todo el espacio muestral, el teorema de Bayes aplica para calcular cualquiera de las probabilidades posteriores $P(A_i \mid B)$ como se muestra a continuación

TEOREMA DE BAYES

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \cdots + P(A_n)P(B \mid A_n)} \tag{4.19}$$

*Si la unión de los eventos es todo el espacio muestral, los eventos son *colectivamente exhaustivos*.

Con las probabilidades previas $P(A_1)$, $P(A_2)$, …, $P(A_n)$ y las probabilidades condicionales adecuadas $P(B \mid A_1)$, $P(B \mid A_2)$, …, $P(B \mid A_n)$, se usa la ecuación (4.19) para calcular la probabilidad posterior de los eventos A_1, A_2, …, A_n

Método tabular

Para realizar los cálculos del teorema de Bayes es útil emplear un método tabular. En la tabla 4.7 se muestra este método aplicado al problema de las piezas de los proveedores. Los cálculos que se muestran ahí se realizan mediante los pasos siguientes.

Paso 1. Se harán las columnas siguientes:
Columna 1: Para los eventos mutuamente excluyentes A_i de los que quiere tener la probabilidad posterior
Columna 2: Para las probabilidades previas $P(A_i)$ de los eventos
Columna 3: Para las probabilidades condicionales $P(B \mid A_i)$ de la nueva información B dado cada evento

Paso 2. En la columna 4 se calculan las probabilidades conjuntas $P(A_i \cap B)$, de cada evento y la nueva información, empleando la ley de la multiplicación. Estas probabilidades conjuntas se encuentran multiplicando las probabilidades previas de la columna 2 por las correspondientes probabilidades condicionales de la columna 3; es decir, $P(A_i \cap B) = P(A_i)P(B \mid A_i)$.

Paso 3. Sume las probabilidades de la columna 4. Esta suma es la probabilidad de la nueva información, $P(B)$. Así, en la tabla 4.7 se ve que la probabilidad de que una pieza sea del proveedor 1 y esté mala es 0.0130 y que la probabilidad de que la pieza sea del proveedor 2 y esté mala es 0.0175. Como éstas son las únicas dos maneras de tener una pieza mala, la suma 0.0130 + 0.0175, que es 0.0305, da la probabilidad de hallar una pieza mala en las piezas recibidas de los dos proveedores.

Paso 4. En la columna 5 se calculan las probabilidades posteriores usando la relación básica de la probabilidad condicional.

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)}$$

Observe que las probabilidades conjuntas $P(A_i \cap B)$ están en la columna 4 y que la probabilidad $P(B)$ es la suma de la columna 4.

TABLA 4.7 MÉTODO TABULAR PARA LOS CÁLCULOS DEL TEOREMA DE BAYES APLICADO AL EJEMPLO DE LOS DOS PROVEEDORES

(1) Eventos A_i	**(2) Probabilidades previas $P(A_i)$**	**(3) Probabilidades condicionales $P(B \mid A_i)$**	**(4) Probabilidades conjuntas $P(A_i \cap B)$**	**(5) Probabilidades posteriores $P(A_i \mid B)$**
A_1	0.65	0.02	0.0130	0.0130/0.0305 = 0.4262
A_2	0.35	0.05	0.0175	0.0175/0.0305 = 0.5738
	1.00		$P(B) = 0.0305$	1.0000

NOTAS Y COMENTARIOS

1. El teorema de Bayes se usa mucho en la toma de decisiones. Las probabilidades previas suelen ser estimaciones subjetivas dadas por la persona que toma las decisiones. Se obtiene información muestral y se usan las probabilidades posteriores para emplearlas en la toma de decisiones.

2. Un evento y su complemento son mutuamente excluyentes y su unión es todo el espacio muestral. Por tanto, el teorema de Bayes siempre se emplea para calcular la probabilidad posterior de un evento y su complemento.

Ejercicios

Métodos



39. Las probabilidades previas de los eventos A_1 y A_2 son $P(A_1) = 0.40$ y $P(A_2) = 0.60$. Sabe también que $P(A_1 \cap A_2) = 0$. Suponga que $P(B \mid A_1) = 0.20$ y $P(B \mid A_2) = 0.05$.
    a. ¿A_1 y A_2 son eventos mutuamente excluyentes? Explique.
    b. Calcule $P(A_1 \cap B)$ y $P(A_2 \cap B)$.
    c. Calcule $P(B)$.
    d. Emplee el teorema de Bayes para calcular $P(A_1 \mid B)$ y $P(A_2 \mid B)$.

40. Las probabilidades previas de los eventos A_1, A_2 y A_3 son $P(A_1) = 0.20$, $P(A_2) = 0.50$ y $P(A_3) = 0.30$. Las probabilidades condicionales del evento B dados los eventos A_1, A_2 y A_3 son $P(B \mid A_1) = 0.50$, $P(B \mid A_2) = 0.40$ y $P(B \mid A_3) = 0.30$.
    a. Calcule $P(B \cap A_1)$, $P(B \cap A_2)$ y $P(B \cap A_3)$.
    b. Emplee el teorema de Bayes, ecuación (4.19), para calcular la probabilidad posterior $P(A_2 \mid B)$.
    c. Use el método tabular para emplear el teorema de Bayes en el cálculo de $P(A_1 \mid B)$, $P(A_2 \mid B)$ y $P(A_3 \mid B)$.

Aplicaciones

41. Una empresa de consultoría presenta una oferta para un gran proyecto de investigación. El director de la firma piensa inicialmente que tiene 50% de posibilidades de obtener el proyecto. Sin embargo, mas tarde, el organismo al que se le hizo la oferta pide más información sobre la oferta. Por experiencia se sabe que en 75% de las ofertas aceptadas y en 40% de las ofertas no aceptadas, este organismo solicita más información.
    a. ¿Cuál es la probabilidad previa de que la oferta sea aceptada (es decir, antes de la solicitud dé más información)?
    b. ¿Cuál es la probabilidad condicional de que se solicite más información dado que la oferta será finalmente aceptada?
    c. Calcule la probabilidad posterior de que la oferta sea aceptada dado que se solicitó más información.

Auto examen

42. Un banco local revisa su política de tarjetas de crédito con objeto de retirar algunas de ellas. En el pasado aproximadamente 5% de los tarjetahabientes incumplieron, dejando al banco sin posibilidad de cobrar el saldo pendiente. De manera que el director estableció una probabilidad previa de 0.05 de que un tarjetahabiente no cumpla. El banco encontró también que la probabilidad de que un cliente que es cumplido no haga un pago mensual es 0.20. Por supuesto la probabilidad de no hacer un pago mensual entre los que incumplen es 1.
    a. Dado que un cliente no hizo el pago de uno o más meses, calcule la probabilidad posterior de que el cliente no cumpla.
    b. El banco deseará retirar sus tarjetas si la probabilidad de que un cliente no cumpla es mayor que 0.20. ¿Debe retirar el banco una tarjeta si el cliente no hace un pago mensual?

43. En los automóviles pequeños el rendimiento de la gasolina es mayor, pero no son tan seguros como los coches grandes. Los automóviles pequeños constituyen 18% de los vehículos en circulación, pero en accidentes con automóviles pequeños se registraron 11 898 victimas mortales en uno de los últimos años (*Reader´s Digest,* mayo de 2000). Suponga que la probabilidad de que un automóvil pequeño tenga un accidente es 0.18. La probabilidad de que en un accidente con un automóvil pequeño haya una víctima mortal es 0.128 y la probabilidad de que haya una víctima mortal si el automóvil no es pequeño es 0.05. Usted se entera de un accidente en el que hubo una víctima mortal. ¿Cuál es la probabilidad de que el accidente lo haya tenido un automóvil pequeño?

44. La American Council of Education informa que en Estados Unidos 47% de los estudiantes que ingresan en la universidad terminan sus estudios en un lapso de cinco años (Associated Press, 6 de mayo de 2002). Suponga que en los registros de terminación de estudios encuentra que 50% de los estudiantes que terminan sus estudios en cinco años son mujeres y 45% de quienes no terminan sus estudios en cinco años son mujeres. Los estudiantes que no terminan sus estudios en cinco años son estudiantes que han abandonado sus estudios o que están por terminarlos.
    a. Sea A_1 =el estudiante termina sus estudios en cinco años
       A_2 = el estudiante no termina sus estudios en cinco
       W = el estudiante es mujer
       Empleando la información dada, dé las probabilidades siguientes: $P(A_1)$, $P(A_2)$, $P(W \mid A_1)$ y $P(W \mid A_2)$.
    b. ¿Cuál es la probabilidad de que una estudiante termine sus estudios en cinco años?
    c. ¿Cuál es la probabilidad de que un estudiante termine sus estudios en cinco años?
    d. Dados los resultados anteriores, ¿cuál es el porcentaje de mujeres y cuál es el porcentaje de hombres que entran en la universidad?

45. En un artículo acerca del crecimiento de las inversiones, la revista *Money* informa que las acciones en medicamentos muestran una poderosa tendencia de largo plazo y ofrecen a los inversionistas potenciales inigualables y duraderas ganancias. La Health Care Financing Administration confirma estas conclusiones con su pronóstico de que para 2010 el consumo de medicamentos llegará a $366 mil millones, cuando en 2000 era de $117 mil millones. Muchas de las personas de 65 años o más necesitan medicamentos. Entre estas personas, 82% necesita medicamentos de manera regular, 55% usa tres o más medicamentos de manera regular y 40% necesita cinco o más medicamentos regularmente. En cambio entre las personas menores de 65 años, 49% usa medicamentos de manera regular, 37% necesita tres o más medicamentos de manera regular y 28% usa cinco o más medicamentos regularmente (*Money*, septiembre de 2001). La Oficina de Censos de Estados Unidos informa que de los 281 421 906 habitantes de Estados Unidos, 34 991 753 son personas de 65 años o mayores (U.S. Census Bureau, *Census 2000*).
    a. Calcule la probabilidad de que en Estados Unidos una persona tenga 65 años o más.
    b. Calcule la probabilidad de que una persona necesite medicamentos de manera regular.
    c. Calcule la probabilidad de que una persona tenga 65 años o más y necesite cinco o más medicamentos.
    d. Dado que una persona usa cinco o más medicamentos, calcule la probabilidad de que tenga 65 años o más.

Resumen

En este capítulo se introdujeron conceptos básicos de probabilidad y se ilustró cómo usar el análisis de probabilidad para obtener información útil para la toma de decisiones. Se describió cómo interpretar la probabilidad como una medida numérica de la posibilidad de que ocurra un evento. Además, se vio que la probabilidad de un evento se puede calcular, ya sea sumando las probabilidades de los resultados experimentales (puntos muestrales) que comprende el evento o usando las relaciones que establecen las leyes de probabilidad de la adición, de la probabilidad condicional y de la multiplicación. En el caso de que se obtenga información adicional, se mostró cómo usar el teorema de Bayes para obtener probabilidades revisadas o posteriores.

Glosario

Probabilidad Medida numérica de la posibilidad de que ocurra un evento.
Experimento Proceso para generar resultados bien definidos.

Espacio muestral Conjunto de todos los resultados experimentales.
Punto muestral Un elemento del espacio muestral. Un punto muestral que representa un resultado experimental.
Diagrama de árbol Representación gráfica que ayuda a visualizar un experimento de pasos múltiples.
Requerimientos básicos en la asignación de probabilidades Dos requerimientos que restringen la manera en que se asignan probabilidades son: 1) Para cada resultado experimental E_i se debe tener $0 \leq P(E_i) \leq 1$; 2) si $E_1, E_2, \ldots, E_n$ son todos los resultados experimentales, se debe tener que $P(E_1) + P(E_2) + \cdots + P(E_n) = 1.0$.
Método clásico Sirve para la asignación de probabilidades, es apropiado cuando todos los resultados experimentales son igualmente posibles.
Método de las frecuencias relativas Útil para la asignación de probabilidades, es conveniente cuando se tienen datos para estimar la proporción de veces que se presentará un resultado experimental si se repite un gran número de veces.
Método subjetivo Método para la asignación de probabilidades basado en un juicio.
Evento Colección de puntos muestrales
Complemento de *A* El evento que consta de todos los puntos muestrales que no están en A.
Diagrama de Venn Una representación gráfica para mostrar de manera simbólica el espacio muestral y las operaciones con eventos en la cual el espacio muestral se representa como un rectángulo y los eventos se representan como círculos dentro del espacio muestral.
Unión de *A* y *B* Evento que contiene todos los puntos muestrales que pertenecen a A o a B o a ambos. La unión se denota $A \cup B$.
Intersección de *A* y *B* Evento que contiene todos los puntos muestrales que pertenecen tanto a A como a B. La intersección se denota $A \cap B$.
Ley de la adición Ley de probabilidad que se usa para calcular la unión de dos eventos. Es $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Si los eventos son mutuamente excluyentes, $P(A \cap B) = 0$; en este caso la ley de la adición se reduce a $P(A \cup B) = P(A) + P(B)$.
Eventos mutuamente excluyentes Eventos que no tienen puntos muestrales en común; es decir, $A \cap B$ es vacío y $P(A \cap B) = 0$.
Probabilidad condicional Probabilidad de un evento dado que otro evento ya ocurrió. La probabilidad condicional de A dado B es $P(A \mid B) = P(A \cap B)/P(B)$.
Probabilidad conjunta La probabilidad de que dos eventos ocurran al mismo tiempo; es decir, la probabilidad de la intersección de dos eventos.
Probabilidad marginal Los valores en los márgenes de una tabla de probabilidad conjunta que dan las probabilidades de cada evento por separado.
Eventos independientes Son dos eventos, A y B, para los que $P(A \mid B) = P(A)$ o $P(B \mid A) = P(B)$; es decir, los eventos no tienen ninguna influencia uno en otro.
Ley de la multiplicación Una ley de probabilidad que se usa para calcular la probabilidad de la intersección de dos eventos. Esto es $P(A \cap B) = P(B)P(A \mid B)$ o $P(A \cap B) = P(A)P(B \mid A)$. Para eventos independientes se reduce a $P(A \cap B) = P(A)P(B)$
Probabilidades previas Estimaciones iniciales de las probabilidades de eventos.
Probabilidades posteriores Probabilidades revisadas de eventos basadas en informaciones adicionales.
Teorema de Bayes Método usado para calcular las probabilidades posteriores.

Fórmulas clave

Regla de conteo para combinaciones

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{4.1}$$

Regla de conteo para permutaciones

$$P_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \quad \textbf{(4.2)}$$

Cálculo de la probabilidad usando el complemento

$$P(A) = 1 - P(A^c) \quad \textbf{(4.5)}$$

Ley de la adición

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \textbf{(4.6)}$$

Probabilidad condicional

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \textbf{(4.7)}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \quad \textbf{(4.8)}$$

Ley de la multiplicación

$$P(A \cap B) = P(B)P(A \mid B) \quad \textbf{(4.11)}$$

$$P(A \cap B) = P(A)P(B \mid A) \quad \textbf{(4.12)}$$

Ley de la multiplicación para eventos independientes

$$P(A \cap B) = P(A)P(B) \quad \textbf{(4.13)}$$

Teorema de Bayes

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \cdots + P(A_n)P(B \mid A_n)} \quad \textbf{(4.19)}$$

Ejercicios complementarios

46. En un sondeo se les pidió a 1035 adultos su opinión respecto a los negocios (*BusinessWeek*, 11 de septiembre de 2000). Una de las preguntas era: "¿Cómo califica usted a las empresas estadounidenses respecto a la calidad de los productos y competitividad a nivel mundial?" Las respuestas fueron: excelentes, 18%; bastante buenas, 50%; regulares, 26%; malas, 5% y no saben o no contestaron 1%.
    a. ¿Cuál es la probabilidad de que un interrogado considere a las empresas estadounidenses bastante buenas o excelentes?
    b. ¿Cuántos de los interrogados consideraron malas a las empresas estadounidenses?
    c. ¿Cuántos de los interrogados dijo no saber o no contestó?
47. Un administrador financiero realiza dos nuevas inversiones, una en la industria del petróleo y otra en bonos municipales. Después de un año cada una de las inversiones se clasificará como buena o no. Considere como un experimento el resultado que se obtiene con estas dos acciones.
    a. ¿Cuántos puntos muestrales hay en este experimento?
    b. Presente un diagrama de árbol y enumere los puntos muestrales.
    c. Sea O = el evento la inversión en la industria del petróleo es buena y M = el evento la inversión en los fondos municipales es buena. Dé los puntos muestrales de O y de M.
    d. Enumere los puntos muestrales de la unión de los eventos $(O \cup M)$.
    e. Cuente los puntos muestrales de la intersección de los eventos $(O \cap M)$.
    f. ¿Son mutuamente excluyentes los eventos O y M? Explique.

48. A principios de 2003, el presidente de Estados Unidos propuso eliminar los impuestos a los dividendos de los accionistas con el argumento de que era un doble impuesto. Las corporaciones pagan impuestos sobre las ganancias que luego son repartidas como dividendos. En un sondeo realizado a 671 estadounidenses, Techno Metrica Market Intelligence halló que 47% estaban a favor de la propuesta, 44% se oponían a ella y 9% no estaban seguros (*Investor's Business Daily,* 13 de enero de 2003). Al analizar las respuestas de acuerdo con la pertenencia a los partidos políticos, se encontró en el sondeo que 29% de los demócratas estaban a favor, 64% de los republicanos estaban a favor y 48% de los independientes estaban a favor.
    a. ¿Cuántos de los encuestados estuvieron a favor de la eliminación de los impuestos a los dividendos?
    b. ¿Cuál es la probabilidad condicional de que una persona esté a favor de la propuesta dado que es demócrata?
    c. ¿Es la afiliación partidaria independiente de que una persona esté a favor de la propuesta?
    d. Si se supone que las respuestas de las personas estuvieron de acuerdo con sus propios intereses, ¿qué grupo se beneficiará más con la aceptación de la propuesta?
49. En un estudio realizado con 31 000 ingresos a hospitales en el estado de Nueva York se encontró que 4% de los ingresados sufrieron daños a causa del tratamiento. Un séptimo de estos daños condujeron a la muerte y un cuarto se debió a negligencia médica. En uno de cada 7.5 casos de negligencia médica se levantó una demanda y en una de cada dos demandas se tuvo que pagar una indemnización.
    a. ¿Cuál es la probabilidad de que una persona que ingresa en un hospital sufra un daño a causa del tratamiento debido a negligencia médica?
    b. ¿Cuál es la probabilidad de que una persona que ingresa en un hospital muera a causa de daños producidos por el tratamiento?
    c. En el caso de daños causado por negligencia médica, ¿cuál es la probabilidad de que la demanda ocasione una indemnización?
50. En una encuesta por teléfono para determinar la opinión de los televidentes respecto a un nuevo programa de televisión se obtuvieron las opiniones siguientes:

Opinión	**Frecuencia**
Malo	4
Regular	8
Bueno	11
Muy bueno	14
Excelente	13

    a. ¿Cuál es la probabilidad de que un televidente tomado aleatoriamente opine que el nuevo programa es bueno o le dé un calificativo mejor.
    b. ¿Cuál es la probabilidad de que un televidente tomado aleatoriamente opine que el nuevo programa es regular o le dé un calificativo inferior?
51. En la siguiente tabulación cruzada se muestra el ingreso familiar de acuerdo con el nivel de estudios del cabeza de familia (*Statistical Abstract of the United States: 2002*).

	Ingreso familiar (en miles de $)					
Nivel de estudios	**Menos de 25**	**25.0–49.9**	**50.0–74.9**	**75.0–99.9**	**100 o más**	**Total**
Preparatoria sin terminar	9 285	4 093	1 589	541	354	15 862
Preparatoria terminada	10 150	9 821	6 050	2 737	2 028	30 786
Estudios universitarios sin terminar	6 011	8 221	5 813	3 215	3 120	26 380
Estudios universitarios terminados	2 138	3 985	3 952	2 698	4 748	17 521
Estudios de posgrado	813	1 497	1 815	1 589	3 765	9 479
Total	28 397	27 617	19 219	10 780	14 015	100 028

a. Elabore una tabla de probabilidad conjunta.
b. ¿Cuál es la probabilidad de que el cabeza de familia no haya terminado la preparatoria?
c. ¿Cuál es la probabilidad de que el cabeza de familia haya terminado la universidad o tenga estudios de posgrado?
d. ¿Cuál es la probabilidad de que si el cabeza de familia terminó la universidad, el ingreso familiar sea $100 000 o más?
e. ¿Cuál es la probabilidad de que el ingreso familiar sea menor a $25 000?
f. ¿Cuál es la probabilidad de que una familia en la que el cabeza de familia terminó la universidad, tenga un ingreso familiar menor a $25 000?
g. ¿El ingreso familiar es independiente del nivel de educación?

52. En un estudio realizado entre los 2010 nuevos estudiantes inscritos a las maestrías de negocios se obtuvieron los datos siguientes.

		Hizo solicitudes en varias universidades	
		Sí	**No**
Grupos de edades	**23 o menos**	207	201
	24–26	299	379
	27–30	185	268
	31–35	66	193
	36 o más	51	169

a. Para un estudiante de maestría tomado en forma aleatoria elabore una tabla de probabilidad conjunta para el experimento que consiste en observar la edad del estudiante y si hizo solicitudes en varias universidades.
b. ¿Cuál es la probabilidad de que un estudiante tomado en forma aleatoria tenga 23 años o menos?
c. ¿Cuál es la probabilidad de que un estudiante tomado en forma aleatoria tenga más de 26 años?
d. ¿Cuál es la probabilidad de que un estudiante tomado en forma aleatoria haya hecho solicitud en varias universidades?

53. Vaya nuevamente a los datos de los nuevos estudiantes inscritos a las maestrías de negocios del ejercicio 52.
a. Dado que una persona hizo solicitudes en varias universidades, ¿cuál es la probabilidad de que tenga entre 24 y 26 años?
b. Ya que una persona tiene 36 años o más, ¿cuál es la probabilidad de que haya hecho solicitudes en varias universidades?
c. ¿Cuál es la probabilidad de que una persona entre 24 y 26 años haya hecho solicitudes en varias universidades?
d. Suponga que la persona sólo hizo solicitud para una universidad. ¿Cuál es la probabilidad de que la persona tenga 31 años o más?
e. ¿La edad y el hacer solicitudes en varias universidades son independientes? Explique.

54. En una encuesta realizada por IBD/TIPPP para obtener información sobre la opinión respecto a las inversiones para el retiro (*Investor´s Business Daily,* 5 de mayo de 2000) se les preguntó a los hombres y mujeres interrogados qué tan importante les parecía que era el nivel de riesgo al elegir una inversión para el retiro. Con los datos obtenidos se elaboró la siguiente tabla de probabilidades conjuntas. “Importante” significa que el interrogado respondió que el nivel de riesgo era importante o muy importante.

	Hombre	**Mujer**	**Total**
Importante	0.22	0.27	0.49
No importante	0.28	0.23	0.51
Total	0.50	0.50	1.00

a. ¿Cuál es la probabilidad de que uno de los interrogados diga que es importante?
b. ¿Cuál es la probabilidad de que una de las mujeres interrogadas diga que es importante?
c. ¿Cuál es la probabilidad de que uno de los hombres interrogados diga que es importante?
d. ¿El nivel de riesgo es independiente del género del interrogado?
e. ¿La opinión de hombres y mujeres difiere respecto al riesgo?

55. Una empresa grande de productos de consumo transmite por televisión publicidad para uno de sus jabones. De acuerdo con una encuesta realizada, se asignaron probabilidades a los eventos siguientes.

$$B = \text{una persona compra el producto}$$
$$S = \text{una persona recuerda haber visto la publicidad}$$
$$B \cap S = \text{una persona compra el producto y recuerda haber visto la publicidad.}$$

Las probabilidades fueron $P(B) = 0.20$, $P(B) = 0.40$ y $P(B \cap S) = 0.12$.

a. ¿Cuál es la probabilidad de que una persona compre el producto dado que recuerda haber visto la publicidad? ¿Ver la publicidad aumenta la probabilidad de que el individuo compre el producto? Si usted tuviera que tomar la decisión, ¿recomendaría que continuara la publicidad (suponiendo que los costos sean razonables)?
b. Si una persona que no compra el producto de la empresa compra el de la competencia. ¿Cuál sería su estimación de la participación de la empresa en el mercado? ¿Esperaría que continuando con la publicidad aumentara la participación de la empresa en el mercado? ¿Por qué sí o por qué no?
c. La empresa probó también otra publicidad y los valores de probabilidad asignados fueron $P(S) = 0.30$, $P(B \cap S) = 0.10$. Dé $P(B \mid S)$ en el caso de esta otra publicidad. ¿Qué publicidad parece tener mejor efecto en la compra de los clientes?

56. Cooper Realty es una empresa inmobiliaria pequeña que se encuentra en Albany, Nueva York y que se especializa en la venta de casas residenciales. Últimamente quiso saber cuál era la posibilidad de que una de las casas que tiene en venta se vendiera en menos de un determinado número de días. Mediante un análisis de 800 casas vendidas por la empresa en los años anteriores se obtuvieron los datos siguientes.

		Días en venta hasta la compra			
		Menos de 30	**31–90**	**Más de 90**	**Total**
Precio pedido inicialmente	**Menos de $150 000**	50	40	10	100
	$150 000–$199 999	20	150	80	250
	$200 000–$250 000	20	280	100	400
	Más de $250 000	10	30	10	50
	Total	100	500	200	800

a. Si A se define como el evento de que la casa esté en venta más de 90 días antes de ser vendida, estime la probabilidad de A.
b. Si B se define como el evento de que el precio inicial sea menor que $150 000, estime la probabilidad de B.
c. ¿Cuál es la probabilidad de $A \cap B$?
d. Suponga que se acaba de firmar un contrato para vender una casa en un precio inicial menor que $150 000, ¿cuál es la probabilidad de que a Cooper Realty le tome menos de 90 días venderla?
e. ¿Los eventos A y B son independientes?

57. Una empresa estudió el número de accidentes ocurridos en su planta de Brownsville, Texas. De acuerdo con información anterior, 6% de los empleados sufrieron accidentes el año pasado. Los directivos creen que un programa especial de seguridad reducirá este año los accidentes a 5%. Se estima además que 15% de los empleados que sufrieron un accidente el año pasado tendrán un accidente este año.
   a. ¿Qué porcentaje de los empleados sufrirá accidentes en los dos años?
   b. ¿Qué porcentaje de los empleados sufrirá por lo menos un accidente en este periodo de dos años?
58. El departamento de recolección de impuestos de Estados Unidos en Dallas, preocupado por las declaraciones de impuestos fraudulentas, cree que la probabilidad de hallar una declaración de impuestos fraudulenta, dado que la declaración contiene deducciones que exceden el estándar, es 0.20. Dado que las deducciones no exceden el estándar, la probabilidad de una declaración fraudulenta disminuye a 0.02. Si 8% de las declaraciones exceden el estándar de deducciones, ¿cuál es la mejor estimación del porcentaje de declaraciones fraudulentas?
59. Una empresa petrolera compra una opción de tierra en Alaska. Los estudios geológicos preliminares asignaron las probabilidades previas siguientes.

$$P(\text{petróleo de alta calidad}) = 0.50$$
$$P(\text{petróleo de calidad media}) = 0.20$$
$$P(\text{que no haya petróleo}) = 0.30$$

   a. ¿Cuál es la probabilidad de hallar petróleo?
   b. Después de 200 pies de perforación en el primer pozo, se toma una prueba de suelo. Las probabilidades de hallar el tipo de suelo identificado en la prueba son las siguientes.

$$P(\text{suelo} \mid \text{petróleo de alta calidad}) = 0.20$$
$$P(\text{suelo} \mid \text{petróleo de calidad media}) = 0.80$$
$$P(\text{suelo} \mid \text{que no haya petróleo}) = 0.20$$

   ¿Cómo debe interpretar la empresa la prueba de suelo? ¿Cuáles son las probabilidades revisadas y cuáles son las nuevas probabilidades de hallar petróleo?
60. Las empresas que hacen negocios por Internet suelen obtener información acerca del visitante de un sitio Web a partir de los sitios visitados previamente. El artículo “Internet Marketing” (*Interfaces,* marzo/abril de 2001) describe cómo los datos sobre el flujo de clics en los sitios Web visitados se usan junto a un modelo de actualización Bayesiano para determinar el género de una persona que visita la Web. ParFore creó un sitio Web para la venta de equipo y ropa para golf. A los directivos de la empresa les gustaría que apareciera una determinada oferta para los visitantes del sexo femenino y otra oferta determinada para los visitantes del sexo masculino. En una muestra de visitas anteriores al sitio Web se sabe que 60% de las personas que visitan el sitio son hombres y 40% mujeres.
   a. ¿Cuál es la probabilidad previa de que el siguiente visitante del sitio Web sea mujer?
   b. Suponga que el actual visitante de ParFore.com visitó previamente el sitio de la Web de Dillard, y que es tres veces más probable que ese sitio sea visitado por mujeres que por hombres. ¿Cuál es la probabilidad revisada de que el visitante actual de ParFore.com sea mujer? ¿Desplegaría la oferta que está dirigida más a hombres o a mujeres?

Caso problema Los jueces del condado de Hamilton

Los jueces del condado de Hamilton llevan miles de casos cada año. En su inmensa mayoría la sentencia queda dictada. Sin embargo, en algunos casos hay apelaciones y algunas apelaciones revocan la sentencia. Kristen DelGuzzi de *The Cincinnati Enquirer* realizó, durante tres años, un estudio sobre los casos llevados por los jueces del condado de Hamilton. En la tabla 4.8 se muestran los resultados de los 182 908 casos llevados por 38 jueces en tribunales de primera instan-

TABLA 4.8 CASOS DESPACHADOS, APELADOS Y REVOCADOS EN LOS TRIBUNALES DEL CONDADO DE HAMILTON



Tribunal de primera instancia

Juez	Casos despachados	Casos apelados	Casos revocados
Fred Cartolano	3 037	137	12
Thomas Crush	3 372	119	10
Patrick Dinkelacker	1 258	44	8
Timothy Hogan	1 954	60	7
Robert Kraft	3 138	127	7
William Mathews	2 264	91	18
William Morrissey	3 032	121	22
Norbert Nadel	2 959	131	20
Arthur Ney Jr.	3 219	125	14
Richard Niehaus	3 353	137	16
Thomas Nurre	3 000	121	6
John O'Connor	2 969	129	12
Robert Ruehlman	3 205	145	18
J. Howard Sundermann	955	60	10
Ann Marie Tracey	3 141	127	13
Ralph Winkler	3 089	88	6
Total	43 945	1762	199

Tribunal de relaciones domésticas

Juez	Casos despachados	Casos apelados	Casos revocados
Penelope Cunningham	2 729	7	1
Patrick Dinkelacker	6 001	19	4
Deborah Gaines	8 799	48	9
Ronald Panioto	12 970	32	3
Total	30 499	106	17

Tribunal municipal

Juez	Casos despachados	Casos apelados	Casos revocados
Mike Allen	6 149	43	4
Nadine Allen	7 812	34	6
Timothy Black	7 954	41	6
David Davis	7 736	43	5
Leslie Isaiah Gaines	5 282	35	13
Karla Grady	5 253	6	0
Deidra Hair	2 532	5	0
Dennis Helmick	7 900	29	5
Timothy Hogan	2 308	13	2
James Patrick Kenney	2 798	6	1
Joseph Luebbers	4 698	25	8
William Mallory	8 277	38	9
Melba Marsh	8 219	34	7
Beth Mattingly	2 971	13	1
Albert Mestemaker	4 975	28	9
Mark Painter	2 239	7	3
Jack Rosen	7 790	41	13
Mark Schweikert	5 403	33	6
David Stockdale	5 371	22	4
John A. West	2 797	4	2
Total	108 464	500	104

cia, tribunales de relaciones domésticas y tribunales municipales. Dos de los jueces (Dinkelacker y Hogan) no prestaron sus servicios en el mismo tribunal durante los tres años completos.

El objetivo del estudio de este periódico fue evaluar el trabajo de los jueces. Las apelaciones suelen ser el resultado de errores cometidos por los jueces, y el periódico deseaba saber qué jueces realizan bien su trabajo y qué jueces cometían demasiados errores. Se le solicita su ayuda para realizar el análisis de datos. Emplee sus conocimientos de probabilidad y de probabilidad condicional para ayudar a la clasificación de los jueces. Podrá analizar también la posibilidad de apelación y de revocación en los casos tratados en los distintos tribunales.

Informe administrativo

Elabore un informe con su clasificación de los jueces. Incluya un análisis de la posibilidad de apelación y de revocación del caso en los tres tribunales. Como mínimo su informe debe contener lo siguiente:

1. La probabilidad de que los casos sean apelados y revocados en los distintos tribunales.
2. La probabilidad, para cada juez, de que un caso sea apelado.
3. La probabilidad, para cada juez, de que un caso sea revocado.
4. La probabilidad, para cada juez, de revocación dada una apelación.
5. Clasifique a los jueces de cada tribunal de mejor a peor. Dé el criterio que usa y proporcione el fundamento que justifique su elección.



CAPÍTULO 5

Distribuciones de probabilidad discreta

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

CITIBANK*

LONG ISLAND CITY, NUEVA YORK

Citibank, una división de Citigroup, proporciona una amplia gama de servicios financieros, que comprende cuentas de cheques y de ahorro, préstamos e hipotecas, seguros y servicios de inversión, todos dentro del marco de una estrategia única llamada Citibanking. Citibanking significa una identidad de marca consistente en todo el mundo, una oferta coherente de productos y servicios de calidad para el cliente. Citibanking permite al cliente disponer de dinero en cualquier momento, en cualquier parte y de la manera que lo desee. Ya sea que el cliente desee ahorrar para el futuro o solicitar un préstamo para hoy, lo puede hacer en Citibank.

Los cajeros automáticos de Citibank, localizados en los Citicard Banking Center (CBC), permiten al cliente hacer todas sus operaciones bancarias en un solo lugar con un simple toque de su dedo, 24 horas al día y 7 días a la semana. Más de 150 operaciones bancarias diferentes, desde depósitos hasta manejo de inversiones, pueden ser realizadas con facilidad. Los cajeros automáticos Citibanking son mucho más que un simple cajero automático y en la actualidad los clientes realizan en ellos 80% de sus transacciones.

Cada Citibank CBC opera como un sistema de espera en línea al que los clientes llegan en forma aleatoria a solicitar el servicio de uno de los cajeros automáticos. Si todos los cajeros automáticos están ocupados, debe esperar en la fila. Con periodicidad realizan estudios acerca de la capacidad de los CBC para determinar los tiempos de espera para el cliente y establecer si son necesarios más cajeros automáticos.

Los datos recolectados por Citibank muestran que la llegada aleatoria de los clientes sigue una distribución de probabilidad conocida como distribución de Poisson. Mediante la distribución de Poisson, Citibank calcula las probabilidades de que llegue un número determinado de clientes a un CBC durante un determinado periodo y decidir cuál es el número de cajeros que necesita. Por ejemplo, sea x la cantidad de clientes que llega en un periodo de un minuto. Suponga que la tasa media de llegadas de clientes a un determinado CBC es dos clientes por minuto, la tabla siguiente da las probabilidades de que llegue un determinado número de clientes por minuto.



Un vanguardista cajero automático de Citibank.
© Jeff Greenberg/Photo Edit.

x	Probabilidad
0	0.1353
1	0.2707
2	0.2707
3	0.1804
4	0.0902
5 o más	0.0527

Las distribuciones de probabilidad discretas como la empleada por Citibank, son el tema de este capítulo. Además de la distribución de Poisson, verá las distribuciones binomial e hipergeométrica; conocerá también cómo emplear estas distribuciones de probabilidad para obtener información de utilidad.



*Los autores agradecen a Stacey Karter, Citibank, por proporcionarnos este artículo para *La estadística en práctica.*

En este capítulo se continúa con el estudio de la probabilidad introduciendo los conceptos de variable aleatoria y distribuciones de probabilidad. El punto sustancial de este capítulo son las distribuciones de probabilidad discreta de tres distribuciones de probabilidad discreta que serán estudiadas son: la binomial, la de Poisson y la hipergeométrica.

5.1 Variables aleatorias

En el capítulo 4 se definió el concepto de experimento con sus correspondientes resultados experimentales. Una variable aleatoria proporciona un medio para describir los resultados experimen-

VARIABLE ALEATORIA

Una **variable aleatoria** es una descripción numérica del resultado de un experimento.

Las variables aleatorias deben tomar valores numéricos.

tales empleando valores numéricos. Las variables aleatorias deben tomar valores numéricos. En efecto, una variable aleatoria asocia un valor numérico a cada uno de los resultados experimentales. El valor numérico de la variable aleatoria depende del resultado del experimento. Una variable aleatoria puede ser *discreta* o *continua*, depende del tipo de valores numéricos que asuma.

Variables aleatorias discretas

A una variable aleatoria que asuma ya sea un número finito de valores o una sucesión infinita de valores tales como 0, 1, 2, . . ., se le llama **variable aleatoria discreta**. Considere, por ejemplo, el siguiente experimento: un contador presenta el examen para certificarse como contador público. El examen tiene cuatro partes. Defina una variable aleatoria x como $x =$ número de partes del examen aprobadas. Ésta es una variable aleatoria discreta porque puede tomar el número finito de valores 0, 1, 2, 3 o 4.

Para tener otro ejemplo de una variable aleatoria discreta considere el experimento de observar los automóviles que llegan a una caseta de peaje. La variable aleatoria que interesa es $x =$ número de automóviles que llega a la caseta de peaje en un día. Los valores que puede tomar la variable aleatoria son los de la secuencia 0, 1, 2, etc. Así, x es una variable aleatoria discreta que toma uno de los valores de esta sucesión infinita.

Aunque los resultados de muchos experimentos se describen mediante valores numéricos, los de otros no. Por ejemplo, en una encuesta se le puede preguntar a una persona si recuerda el mensaje de un comercial de televisión. Este experimento tiene dos resultados: que la persona no recuerda el mensaje y que la persona recuerda el mensaje. Sin embargo, estos resultados se describen numéricamente definiendo una variable aleatoria x como sigue: sea $x = 0$ si la persona no recuerda el mensaje y sea $x = 1$ si la persona recuerda el mensaje. Los valores numéricos de esta variable son arbitrarios (podría haber usado 5 y 10), pero son aceptables de acuerdo con la definición de una variable aleatoria, es decir, x es una variable aleatoria porque proporciona una descripción numérica de los resultados del experimento.

En la tabla 5.1 aparecen algunos otros ejemplos de variables aleatorias discretas. Observe que en cada ejemplo la variable aleatoria discreta asume un número finito de valores o asume los valores de una secuencia infinita como 0, 1, 2, ... Este tipo de variables aleatorias discretas se estudia con detalle en este capítulo.

TABLA 5.1 EJEMPLOS DE VARIABLES ALEATORIAS DISCRETAS

Experimento	Variable aleatoria (x)	Valores posibles para la variable aleatoria
Llamar a cinco clientes	Número de clientes que hacen un pedido	0, 1, 2, 3, 4, 5
Inspeccionar un envío de 50 radios	Número de radios que tienen algún defecto	0, 1, 2, · · · , 49, 50
Hacerse cargo de un restaurante durante un día	Número de clientes	0, 1, 2, 3, · · ·
Vender un automóvil	Sexo del cliente	0 si es hombre; 1 si es mujer

Variables aleatorias continuas

A una variable que puede tomar cualquier valor numérico dentro de un intervalo o colección de intervalos se le llama **variable aleatoria continua**. Los resultados experimentales basados en escalas de medición tales como tiempo, peso, distancia y temperatura pueden ser descritos por variables aleatorias continuas. Considere, por ejemplo, el experimento de observar las llamadas telefónicas que llegan a la oficina de atención de una importante empresa de seguros. La variable aleatoria que interesa es $x =$ tiempo en minutos entre dos llamadas consecutivas. Esta variable aleatoria puede tomar cualquier valor en el intervalo $x \geq 0$. En efecto, x puede tomar un número infinito de valores, entre los que se encuentran valores como 1.26 minutos, 2.751 minutos, 4.3333 minutos, etc. Otro ejemplo, considere el tramo de 90 millas de una carretera entre Atlanta y Georgia. Para el servicio de ambulancia de emergencia en Atlanta, la variable aleatoria x es $x =$ número de millas hasta el punto en que se localiza el siguiente accidente de tráfico en este tramo de la carretera. En este caso, x es una variable aleatoria continua que toma cualquier valor en el intervalo $0 \leq x \leq 90$. En la tabla 5.2 aparecen otros ejemplos de variables aleatorias continuas. Observe que cada ejemplo describe una variable aleatoria que toma cualquier valor dentro de un intervalo de valores. Las variables aleatorias continuas y sus distribuciones de probabilidad serán tema del capítulo 6.

TABLA 5.2 EJEMPLOS DE VARIABLES ALEATORIAS CONTINUAS

Experimento	Variable aleatoria (x)	Valores posibles para la variable aleatoria
Operar un banco	Tiempo en minutos entre la llegada de los clientes	$x \geq 0$
Llenar una lata de refresco (máx. 12.1 onzas)	Cantidad de onzas	$0 \leq x \leq 12.1$
Construir una biblioteca	Porcentaje del proyecto terminado en seis meses	$0 \leq x \leq 100$
Probar un proceso químico nuevo	Temperatura a la que tiene lugar la reacción deseada (mín. 150°F; máx. 212°F)	$150 \leq x \leq 212$

NOTAS Y COMENTARIOS

Un modo de determinar si una variable aleatoria es discreta o continua es imaginar los valores de la variable aleatoria como puntos sobre un segmento de recta. Elegir dos puntos que representen valores de la variable aleatoria. Si todo el segmento de recta entre esos dos puntos representa también valores posibles para la variable aleatoria, entonces la variable aleatoria es continua.

Ejercicios

Métodos



1. Considere el experimento que consiste en lanzar una moneda dos veces.
   a. Enumere los resultados experimentales.
   b. Defina una variable aleatoria que represente el número de caras en los dos lanzamientos.
   c. Dé el valor que la variable aleatoria tomará en cada uno de los resultados experimentales.
   d. ¿Es una variable aleatoria discreta o continua?

2. Considere el experimento que consiste en un empleado que arma un producto.
   a. Defina la variable aleatoria que represente el tiempo en minutos requerido para armar el producto.
   b. ¿Qué valores toma la variable aleatoria?
   c. ¿Es una variable aleatoria discreta o continua?

Aplicaciones



3. Tres estudiantes agendan entrevistas para un empleo de verano en el Brookwood Institute. En cada caso el resultado de la entrevista será una oferta de trabajo o ninguna oferta. Los resultados experimentales se definen en términos de los resultados de las tres entrevistas.
   a. Enumere los resultados experimentales.
   b. Defina una variable aleatoria que represente el número de ofertas de trabajo. ¿Es una variable aleatoria continua?
   c. Dé el valor de la variable aleatoria que corresponde a cada uno de los resultados experimentales.
4. Suponga que conoce la tasa hipotecaria de 12 instituciones de préstamo. La variable aleatoria que interesa es el número de las instituciones de préstamo en este grupo que ofrecen una tasa fija a 30 años de 8.5% o menos. ¿Qué valores toma esta variable aleatoria?
5. Para realizar cierto análisis de sangre, los técnicos laboratoristas tienen que llevar a cabo dos procedimientos. En el primero requieren uno o dos pasos y en el segundo requieren uno, dos o tres pasos.
   a. Enumere los resultados experimentales correspondientes a este análisis de sangre.
   b. Si la variable aleatoria que interesa es el número de pasos requeridos en todo el análisis (los dos procedimientos), dé los valores que toma la variable aleatoria en cada uno de los resultados experimentales.
6. A continuación se da una serie de experimentos y su variable aleatoria correspondiente. En cada caso determine qué valores toma la variable aleatoria y diga si se trata de una variable aleatoria discreta o continua.

Experimento	**Variable aleatoria (x)**
a. Hacer un examen con 20 preguntas	Número de preguntas contestadas correctamente
b. Observar los automóviles que llegan a una caseta de peaje en 1 hora	Número de automóviles que llegan a la caseta de peaje
c. Revisar 50 declaraciones de impuestos	Número de declaraciones que tienen algún error
d. Observar trabajar a un empleado	Número de horas no productivas en una jornada de 8 horas
e. Pesar un envío	Número de libras

5.2 Distribuciones de probabilidad discreta

La **disribución de probabilidad** de una variable aleatoria describe cómo se distribuyen las probabilidades entre los valores de la variable aleatoria. En el caso de una variable aleatoria discreta x, la distribución de probabilidad está definida por una **función de probabilidad**, denotada por $f(x)$. La función de probabilidad da la probabilidad de cada valor de la variable aleatoria.

Como ejemplo de una variable aleatoria discreta y de su distribución de probabilidad, considere las ventas de automóviles en DiCarlo Motors en Saratoga, Nueva York. Durante los últimos 300 días de operación, los datos de ventas muestran que hubo 57 días en los que no se vendió ningún automóvil, 117 días en los que se vendió 1 automóvil, 72 días en los que se vendieron 2 automóviles, 42 días en los que se vendieron 3 automóviles, 12 días en los que se vendieron 4 automóviles y 3 días en los que se vendieron 5 automóviles. Suponga que considera el experimento

de seleccionar un día de operación en DiCarlo Motors y se define la variable aleatoria de interés como x = número de automóviles vendidos en un día. De acuerdo con datos del pasado, se sabe que x es una variable aleatoria discreta que puede tomar los valores 0, 1, 2, 3, 4 o 5. En la notación de funciones de probabilidad $f(0)$ da la probabilidad de vender 0 automóviles, $f(1)$ da la probabilidad de vender 1 automóvil, y así en lo sucesivo. Como los datos del pasado indican que en 54 de 300 días se vendieron 0 automóviles, a $f(0)$ se le asigna el valor $54/300 = 0.18$, lo que significa que la probabilidad de que se vendan 0 automóviles en un día es 0.18. De manera similar, como en 117 de los 300 días se vendió un automóvil, a $f(1)$ se le asigna el valor $117/300 = 0.39$, que significa que la probabilidad de que se venda exactamente 1 automóvil en un día es 0.39. Continuando de esta manera con los demás valores de la variable aleatoria, se obtienen los valores de $f(2)$, $f(3)$, $f(4)$ y $f(5)$, valores que se muestran en la tabla 5.3, que es la distribución de probabilidad para el número de automóviles vendidos en un día en DiCarlo Motors.

Una ventaja importante de definir una variable aleatoria y su correspondiente distribución de probabilidad es que una vez que se conoce la distribución de probabilidad, es relativamente fácil determinar la probabilidad de diversos eventos que pueden ser útiles para tomar decisiones. Por ejemplo, empleando la distribución de probabilidad de DiCarlo Motors, tabla 5.3, se observa que el número de automóviles que es más probable vender en un día es 1, ya que es $f(1) = 0.39$. Además se observa que la probabilidad de vender tres o más automóviles en un día es $f(3) + f(4) + f(5) = 0.14 + 0.04 + 0.01 = 0.19$. Estas probabilidades, junto con otras que pueden interesar para tomar decisiones, proporcionan información que sirve de ayuda al encargado de la toma de decisiones para entender la venta de automóviles en DiCarlo Motors.

Al elaborar una función de probabilidad para una variable aleatoria discreta, deben satisfacerse las dos condiciones siguientes.

Estas condiciones son análogas a los dos requerimientos básicos, presentados en el capítulo 4, para asignar probabilidades a los resultados experimentales.

CONDICIONES REQUERIDAS PARA UNA FUNCIÓN DE PROBABILIDAD DISCRETA

$$f(x) \geq 0 \quad \textbf{(5.1)}$$

$$\Sigma f(x) = 1 \quad \textbf{(5.2)}$$

En la tabla 5.3 se observa que las probabilidades de la variable aleatoria x satisfacen la ecuación (5.1); para todos los valores de x, $f(x)$ es mayor o igual que 0; además, como estas probabilidades suman 1, también se satisface la ecuación (5.2). Por tanto, la función de probabilidad de DiCarlo Motors es una función de probabilidad discreta válida.

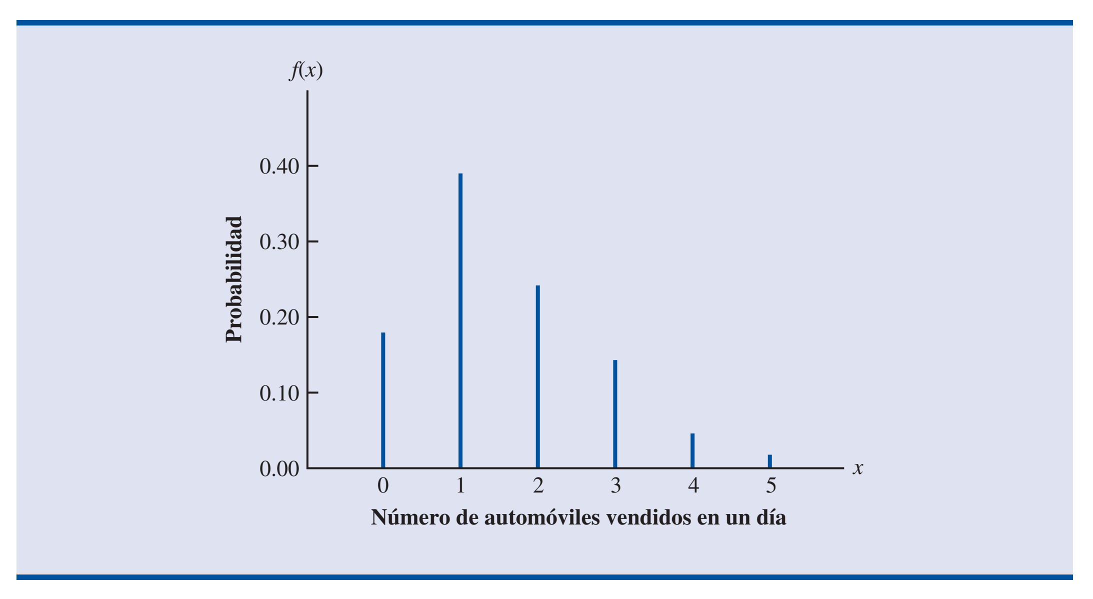
Las distribuciones de probabilidad también se representan gráficamente. En la figura 5.1, en el eje horizontal aparecen los valores de la variable aleatoria x para el caso de DiCarlo Motors y en el eje vertical aparecen las probabilidades correspondientes a estos valores.

Además de tablas y gráficas, para describir las funciones de probabilidad se suele usar una fórmula que da el valor de la función de probabilidad, $f(x)$, para cada valor x. El ejemplo más sencillo

TABLA 5.3 DISTRIBUCIÓN DE PROBABILIDAD PARA EL NÚMERO DE AUTOMÓVILES VENDIDOS EN UN DÍA EN DICARLO MOTORS

x	$f(x)$
0	0.18
1	0.39
2	0.24
3	0.14
4	0.04
5	0.01
Total	1.00

FIGURA 5.1 REPRESENTACIÓN GRÁFICA DE LA DISTRIBUCIÓN DE PROBABILIDAD DEL NÚMERO DE AUTOMÓVILES VENDIDOS EN UN DÍA EN DICARLO MOTORS



de una distribución de probabilidad discreta dada mediante una fórmula es la **distribución de probabilidad uniforme discreta**. Su función de probabilidad está definida por la ecuación (5.3).

FUNCIÓN DE PROBABILIDAD UNIFORME DISCRETA

$$f(x) = 1/n \qquad \textbf{(5.3)}$$

donde

n = número de valores que puede tomar la variable aleatoria.

Por ejemplo, si en el experimento que consiste en lanzar un dado se define una variable aleatoria x como el número de puntos en la cara del dado que cae hacia arriba. En este experimento la variable aleatoria toma $n = 6$ valores; $x = 1, 2, 3, 4, 5, 6$. Por tanto, la función de probabilidad de esta variable aleatoria uniforme discreta es

$$f(x) = 1/6 \qquad x = 1, 2, 3, 4, 5, 6$$

Los valores de la variable aleatoria con sus probabilidades correspondientes se presentan a continuación.

x	$f(x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Otro ejemplo, la variable aleatoria x tiene la siguiente distribución de probabilidad discreta.

x	$f(x)$
1	1/10
2	2/10
3	3/10
4	4/10

Esta distribución de probabilidad se define mediante la fórmula

$$f(x) = \frac{x}{10} \quad \text{para } x = 1, 2, 3 \text{ o } 4$$

Si evalúa $f(x)$ para un valor determinado de la variable aleatoria obtiene la probabilidad correspondiente. Por ejemplo, con la función de probabilidad dada arriba se ve que $f(2) = 2/10$ da la probabilidad de que la variable aleatoria tome el valor 2.

Las funciones de probabilidad discreta más empleadas suelen especificarse mediante fórmulas. Tres casos importantes son las distribuciones binomial, de Poisson e hipergeométrica; estas distribuciones se estudian más adelante en este capítulo

Ejercicios

Métodos



7. A continuación se presenta la distribución de probabilidad de una variable aleatoria x.

x	$f(x)$
20	0.20
25	0.15
30	0.25
35	0.40

   a. ¿Es válida esta distribución de probabilidad?
   b. ¿Cuál es la probabilidad de que $x = 30$?
   c. ¿Cuál es la probabilidad de que x sea menor o igual que 25?
   d. ¿Cuál es la probabilidad de que x sea mayor que 30?

Aplicaciones



8. Los datos siguientes se obtuvieron contando el número de salas de operaciones de un hospital que fueron usadas en un periodo de 20 días. Tres de estos 20 días sólo se usó una sala de operaciones, cinco de estos 20 días se usaron dos, ocho de estos 20 días se usaron tres salas de operaciones y cuatro de estos 20 días se usaron las cuatro salas de operaciones del hospital.
   a. Use el método de las frecuencias relativas para elaborar una distribución de probabilidad para el número de salas de operaciones usadas en un día.
   b. Elabore una gráfica a partir de la distribución de probabilidad.
   c. Muestre que la distribución de probabilidad elaborada satisface las condiciones requeridas para una distribución de probabilidad.

9. En Estados Unidos 38% de los niños de cuarto grado no pueden leer un libro adecuado a su edad. La tabla siguiente muestra, de acuerdo con las edades, el número de niños que tienen problemas de lectura. La mayoría de estos niños tienen problemas de lectura que debieron ser detectados y corregidos antes del tercer grado.

Edad	Número de niños
6	37 369
7	87 436
8	160 840
9	239 719
10	286 719
11	306 533
12	310 787
13	302 604
14	289 168

Si desea tomar una muestra de niños que tienen problemas de lectura para que participen en un programa que mejora las habilidades de lectura. Sea x la variable aleatoria que indica la edad de un niño tomado en forma aleatoria.

a. Con estos datos elabore una distribución de probabilidad para x. Especifique los valores de la variable aleatoria y los correspondientes valores de la función de probabilidad $f(x)$.
b. Trace la gráfica de esta distribución de probabilidad.
c. Muestre que la distribución de probabilidad satisface las ecuaciones (5.1) y (5.2).

10. En la tabla 5.4 se muestra la distribución de frecuencias porcentuales para las puntuaciones dadas a la satisfacción con el trabajo por una muestra de directivos en sistemas de información de nivel alto y de nivel medio. Las puntuaciones van de 1 (muy insatisfecho) a 5 (muy satisfecho).

TABLA 5.4 DISTRIBUCIÓN DE FRECUENCIA PORCENTUAL DE LAS PUNTUACIONES DADAS POR DIRECTIVOS DE NIVEL ALTO Y DE NIVEL MEDIO A LA SATISFACCIÓN CON EL TRABAJO

Puntuación de la satisfacción con el trabajo	Directivos de alto nivel	Directivos de nivel medio
1	5	4
2	9	10
3	3	12
4	42	46
5	41	28

a. Elabore una distribución de probabilidad con las puntuaciones dadas a la satisfacción con el trabajo por los directivos de nivel alto.
b. Elabore una distribución de probabilidad con las puntuaciones dadas a la satisfacción con el trabajo por los directivos de nivel medio.
c. ¿Cuál es la probabilidad de que un ejecutivo de nivel alto dé una puntuación de 4 o 5 a su satisfacción con el trabajo?
d. ¿Cuál es la probabilidad de que un ejecutivo de nivel medio esté muy satisfecho?
e. Haga una comparación entre la satisfacción con el trabajo de los ejecutivos de nivel alto y la que tienen los ejecutivos de nivel medio.

11. Un técnico da servicio a máquinas franqueadoras de empresas en el área de Phoenix. El servicio puede durar 1, 2, 3 o 4 horas dependiendo del tipo de falla. Los distintos tipos de fallas se presentan aproximadamente con la misma frecuencia.

a. Elabore una distribución de probabilidad de las duraciones de los servicios.
b. Elabore una gráfica de la distribución de probabilidad.
c. Muestre que la distribución de probabilidad que ha elaborado satisface las condiciones requeridas para ser una distribución de probabilidad discreta.
d. ¿Cuál es la probabilidad de que un servicio dure tres horas?
e. Acaba de llegar una solicitud de servicio y no se sabe cuál es el tipo de falla. Son las 3:00 p.m. y los técnicos de servicio salen a las 5:00 de la tarde. ¿Cuál es la probabilidad de que el técnico de servicio tenga que trabajar horas extras para reparar la máquina hoy?

12. El jefe del departamento de admisión de una universidad calcula subjetivamente una distribución de probabilidad para x, el número de estudiantes que ingresarán en la universidad. A continuación se presenta esta distribución de probabilidad.

x	$f(x)$
1000	0.15
1100	0.20
1200	0.30
1300	0.25
1400	0.10

a. ¿Es válida esta distribución de probabilidad? Explique.
b. ¿Cuál es la probabilidad de que ingresen 1200 o menos estudiantes? Explique.

13. Un psicólogo encuentra que el número de sesiones necesarias para ganarse la confianza de un paciente es 1, 2 o 3. Sea x la variable aleatoria que representa el número de sesiones necesarias para ganarse la confianza de un paciente. Se ha propuesto la función de probabilidad siguiente.

$$f(x) = \frac{x}{6} \qquad \text{para } x = 1, 2 \text{ o } 3$$

a. ¿Es válida esta función de probabilidad? Explique.
b. ¿Cuál es la probabilidad de que se necesiten exactamente 2 sesiones para ganarse la confianza del paciente?
c. ¿De que se necesiten por lo menos 2 sesiones para ganarse la confianza del paciente?

14. La tabla siguiente es una distribución parcial de probabilidades para las ganancias proyectadas de MRA Company (x ganancias en miles de dólares) durante el primer año de operación (los valores negativos indican pérdida).

x	$f(x)$
−100	0.10
0	0.20
50	0.30
100	0.25
150	0.10
200	

a. ¿Cuál es el valor adecuado para $f(200)$? ¿Qué interpretación le da a este valor?
b. ¿Cuál es la probabilidad de que la empresa sea rentable?
c. ¿Cuál es la probabilidad de que la empresa gane por lo menos $100 000?



Valor esperado y varianzas

Valor esperado

El **valor esperado**, o media, de una variable aleatoria es una medida de la localización central de la variable aleatoria. A continuación se da la fórmula para obtener el valor esperado de una variable aleatoria x.

El valor esperado es un promedio ponderado de los valores que toma la variable aleatoria. Los pesos son las probabilidades.

VALOR ESPERADO DE UNA VARIABLE ALEATORIA DISCRETA

$$E(x) = \mu = \Sigma xf(x) \quad \textbf{(5.4)}$$

Las dos notaciones $E(x)$ y μ se usan para denotar el valor esperado de una variable aleatoria x.

La ecuación (5.4) indica que para calcular el valor esperado de una variable aleatoria discreta se multiplica cada valor de la variable aleatoria por su probabilidad correspondiente $f(x)$ y después se suman estos productos. Usando el ejemplo de la sección 5.2 sobre las ventas de automóviles en DiCarlo Motors, en la tabla 5.5 se muestra cómo se calcula el valor esperado del número de automóviles vendidos en un día. La suma de las entradas en la columna $xf(x)$ indica que el valor esperado es 1.50 automóviles por día. Por tanto, aunque se sabe que en un día las ventas pueden ser de 0, 1, 2, 3, 4 o 5 automóviles, DiCarlo prevé que a la larga se venderán 1.50 automóviles por día. Si en un mes hay 30 días de operación, el valor esperado, 1.50, se emplea para pronosticar que las ventas promedio mensuales serán de 30(1.5) = 45 automóviles.

El valor esperado no tiene que ser un valor que pueda tomar la variable aleatoria.

Varianza

Aunque el valor esperado proporciona el valor medio de una variable aleatoria, también suele ser necesaria una medida de la variabilidad o dispersión. Así como en el capítulo 3 se usó la varianza para resumir la variabilidad de los datos, ahora se usa la **varianza** para resumir la variabilidad en los valores de la variable aleatoria. A continuación se da la fórmula para calcular la

La varianza es un promedio ponderado de los cuadrados de las desviaciones de una variable aleatoria de su media. Los pesos son las probabilidades.

VARIANZA DE UNA VARIABLE ALEATORIA DISCRETA

$$\text{Var}(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x) \quad \textbf{(5.5)}$$

TABLA 5.5 CÁLCULO DEL VALOR ESPERADO PARA EL NÚMERO DE AUTOS QUE SE VENDEN EN UN DÍA EN DICARLO MOTORS

x	$f(x)$	$xf(x)$
0	0.18	0(.18) = 0.00
1	0.39	1(.39) = 0.39
2	0.24	2(.24) = 0.48
3	0.14	3(.14) = 0.42
4	0.04	4(.04) = 0.16
5	0.01	5(.01) = 0.05
		1.50

$E(x) = \mu = \Sigma xf(x)$

TABLA 5.6 CÁLCULO DE LA VARIANZA PARA EL NÚMERO DE AUTOS QUE SE VENDEN EN UN DÍA EN DICARLO MOTORS

x	$x - \mu$	$(x - \mu)^2$	$f(x)$	$(x - \mu)^2 f(x)$
0	$0 - 1.50 = -1.50$	2.25	0.18	$2.25(0.18) = 0.4050$
1	$1 - 1.50 = -0.50$	0.25	0.39	$0.25(0.39) = 0.0975$
2	$2 - 1.50 = 0.50$	0.25	0.24	$0.25(0.24) = 0.0600$
3	$3 - 1.50 = 1.50$	2.25	0.14	$2.25(0.14) = 0.3150$
4	$4 - 1.50 = 2.50$	6.25	0.04	$6.25(0.04) = 0.2500$
5	$5 - 1.50 = 3.50$	12.25	0.01	$12.25(0.01) = 0.1225$
				1.2500

$\sigma^2 = \Sigma(x - \mu)^2 f(x)$

varianza de una variable aleatoria. Como indica la ecuación (5.5), un parte esencial de la fórmula de la varianza es la desviación $x - \mu$, la cual mide qué tan alejado del valor esperado, o media μ, se encuentra un valor determinado de la variable aleatoria. Para calcular la varianza de una variable aleatoria, estas desviaciones se elevan al cuadrado y después se ponderan con el correspondiente valor de la función de probabilidad. A la suma de estas desviaciones al cuadrado, ponderadas, se le conoce como *varianza*. Para denotar la varianza de una variable aleatoria se usan las notaciones Var(x) y σ^2.

En la tabla 5.6 aparece en forma resumida el cálculo de la varianza de la distribución de probabilidad del número de automóviles vendidos en un día en DiCarlo Motors. Como ve, la varianza es 1.25. La **desviación estándar**, σ, se define como la raíz cuadrada positiva de la varianza. Por tanto, la desviación estándar del número de automóviles vendidos en un día es

$$\sigma = \sqrt{1.25} = 1.118$$

La desviación estándar se mide en las mismas unidades que la variable aleatoria ($\sigma = 1.1180$ automóviles) y por tanto suele preferirse para describir la variabilidad de una variable aleatoria. La varianza σ^2 se mide en unidades al cuadrado por lo que es más difícil de interpretar.

Ejercicios

Métodos

15. La tabla siguiente muestra la distribución de probabilidad de una variable aleatoria x.

x	$f(x)$
3	0.25
6	0.50
9	0.25

a. Calcule $E(x)$, el valor esperado de x.
b. Calcule σ^2, la varianza de x.
c. Calcule σ, la desviación estándar de x.



16. La tabla siguiente muestra la distribución de probabilidad de una variable aleatoria y.

y	$f(y)$
2	0.20
4	0.30
7	0.40
8	0.10

a. Calcule $E(y)$.
b. Calcule $\text{Var}(y)$ y σ.

Aplicaciones

17. Una ambulancia de voluntarios realiza de 0 a 5 servicios por día. A continuación se presenta la distribución de probabilidad de los servicios por día.

Número de servicios	Probabilidad	Número de servicios	Probabilidad
0	0.10	3	0.20
1	0.15	4	0.15
2	0.30	5	0.10

a. ¿Cuál es el valor esperado del número de servicios?
b. ¿Cuál es la varianza del número de servicios? ¿Cuál es la desviación estándar?



18. Los datos siguientes son el número de recámaras en casas rentadas y en casas propias en ciudades centrales de Estados Unidos (www.census.gov, 31 de marzo de 2003).

	Número de casas (en miles)	
Recámaras	**Rentadas**	**Propias**
0	547	23
1	5012	541
2	6100	3832
3	2644	8690
4 o más	557	3783

a. Defina una variable aleatoria x = número de recámaras en casas rentadas y elabore una distribución de probabilidad para esta variable. ($x = 4$ representará 4 recámaras o más.)
b. Calcule el valor esperado y la varianza del número de recámaras en casas rentadas.
c. Defina una variable aleatoria y = número de recámaras en casas propias y elabore una distribución de probabilidad para esta variable. ($y = 4$ representará 4 recámaras o más.)
d. Calcule el valor esperado y la varianza del número de recámaras en casas propias.
e. ¿Qué observaciones resultan al comparar el número de recámaras en casas rentadas y en casas propias?

19. La National Basketball Association (NBA) lleva diversas estadísticas de cada equipo. Dos se refieren al porcentaje de tiros de campo hechos por un equipo y el porcentaje de tiros de tres puntos hechos por un equipo. En parte de la temporada del 2004, el registro de tiros de los 29 equipos de la NBA indicaba que la probabilidad de anotar dos puntos en un tiro de campo era 0.44, y que la probabilidad de anotar tres puntos en un tiro de tres puntos era 0.34 (www.nba.com, 3 de enero de 2004).

a. ¿Cuál es el valor esperado para un tiro de dos puntos de estos equipos?
b. ¿Cuál es el valor esperado para un tiro de tres puntos de estos equipos?
c. Si la probabilidad de hacer un tiro de dos puntos es mayor que la probabilidad de hacer uno de tres puntos, ¿por qué los entrenadores permiten a algunos jugadores hacer un tiro de tres puntos si tienen oportunidad? Use el valor esperado para explicar su respuesta.

20. A continuación se presenta la distribución de probabilidad para los daños pagados por una empresa de seguros para automóviles, en seguros contra choques.

Pago	Probabilidad
0	0.85
500	0.04
1 000	0.04
3 000	0.03
5 000	0.02
8 000	0.01
10 000	0.01

a. Use el pago esperado para determinar la prima en el seguro de choques que le permitirá a la empresa cubrir los gastos.
b. La empresa de seguros cobra una tasa anual de $520 por la cobertura de choques. ¿Cuál es el valor esperado de un seguro de choques para un asegurado? (*Indicación*: son los pagos esperados de la empresa menos el costo de cobertura.) ¿Por qué compran los asegurados un seguro de choques con este valor esperado?

21. La siguiente distribución de probabilidad sobre puntuaciones dadas a la satisfacción con el trabajo por una muestra de directivos de alto nivel y de nivel medio en sistemas de la información va desde 1 (muy insatisfecho) hasta 5 (muy satisfecho).

	Probabilidad	
Puntuación de la satisfacción con el trabajo	**Directivo de nivel alto**	**Directivo de nivel medio**
1	0.05	0.04
2	0.09	0.10
3	0.03	0.12
4	0.42	0.46
5	0.41	0.28

a. ¿Cuál es el valor esperado en las puntuaciones dadas a la satisfacción con el trabajo por los ejecutivos de nivel alto?
b. ¿Cuál es el valor esperado en las puntuaciones dadas a la satisfacción con el trabajo por los directivos de nivel medio?
c. Calcule la varianza de las puntuaciones dadas a la satisfacción con el trabajo por los directivos de nivel medio.
d. Calcule la desviación estándar de las puntuaciones dadas a la satisfacción con el trabajo en las dos distribuciones de probabilidad.
e. Compare la satisfacción con el trabajo de los directivos de alto nivel con la que tienen los directivos de nivel medio.

22. La demanda de un producto de una empresa varía enormemente de mes a mes. La distribución de probabilidad que se presenta en la tabla siguiente, basada en los datos de los dos últimos años, muestra la demanda mensual de la empresa.

Demanda unitaria	Probabilidad
300	0.20
400	0.30
500	0.35
600	0.15

a. Si la empresa basa las órdenes mensuales en el valor esperado de la demanda mensual, ¿cuál será la cantidad ordenada mensualmente por la empresa para este producto?
b. Suponga que cada unidad demandada genera $70 de ganancia y que cada unidad ordenada cuesta $50. ¿Cuánto ganará o perderá la empresa en un mes si coloca una orden con base en su respuesta al inciso a y la demanda real de este artículo es de 300 unidades?

23. El estudio 2002 New York City Housing and Vacancy Survey indicó que había 59 324 viviendas con renta controlada y 236 263 unidades con renta estabilizada construidas en 1947 o después. A continuación se da la distribución de probabilidad para el número de personas que viven en estas unidades (www.census.gov, 12 de enero de 2004).

Número de personas	Renta controlada	Renta estabilizada
1	0.61	0.41
2	0.27	0.30
3	0.07	0.14
4	0.04	0.11
5	0.01	0.03
6	0.00	0.01

a. ¿Cuál es el valor esperado para el número de personas que viven en cada tipo de unidad?
b. ¿Cuál es la varianza para el número de personas que viven en cada tipo de unidad?
c. Haga comparaciones entre el número de personas que viven en una unidad de renta controlada y el número de personas que viven en una unidad de renta estabilizada.

24. J. R. Ryland Computer Company está considerando hacer una expansión a la fábrica para empezar a producir una nueva computadora. El presidente de la empresa debe determinar si hacer un proyecto de expansión a mediana gran escala. La demanda del producto nuevo es incierta, la cual, para los fines de planeación puede ser demanda pequeña, mediana o grande. Las probabilidades estimadas para la demanda son 0.20, 0.50 y 0.30, respectivamente. Con x y y representando ganancia anual en miles de dólares, los encargados de planeación en la empresa elaboraron el siguiente pronóstico de ganancias para los proyectos de expansión a mediana y gran escala.

		Ganancia con la expansión a mediana escala		Ganancia con la expansión a gran escala	
		x	$f(x)$	y	$f(y)$
Demanda	Baja	50	0.20	0	0.20
	Mediana	150	0.50	100	0.50
	Alta	200	0.30	300	0.30

a. Calcule el valor esperado de las ganancias correspondientes a las dos alternativas de expansión. ¿Cuál de las decisiones se prefiere para el objetivo de maximizar la ganancia esperada?
b. Calcule la varianza de las ganancias correspondientes a las dos alternativas de expansión. ¿Cuál de las decisiones se prefiere para el objetivo de minimizar el riesgo o la incertidumbre?

5.4 Distribución de probabilidad binomial

La distribución de probabilidad binomial es una distribución de probabilidad que tiene muchas aplicaciones. Está relacionada con un experimento de pasos múltiples al que se le llama experimento binomial.

Un experimento binomial

Un **experimento binomial** tiene las cuatro propiedades siguientes.

PROPIEDADES DE UN EXPERIMENTO BINOMIAL

1. El experimento consiste en una serie de n ensayos idénticos.
2. En cada ensayo hay dos resultados posibles. A uno de estos resultados se le llama *éxito* y al otro se le llama *fracaso*.
3. La probabilidad de éxito, que se denota p, no cambia de un ensayo a otro. Por ende, la probabilidad de fracaso, que se denota $1 - p$, tampoco cambia de un ensayo a otro.
4. Los ensayos son independientes.

Jacob Bernoulli (1654-1705), el primero de la familia Bernoulli de matemáticos suizos, publicó un tratado sobre probabilidad que contenía la teoría de las permutaciones y de las combinaciones, así como el teorema del binomio.

Si se presentan las propiedades 2, 3 y 4, se dice que los ensayos son generados por un proceso de Bernoulli. Si, además, se presenta la propiedad 1, se trata de un experimento binomial. En la figura 5.2 se presenta una sucesión de éxitos y fracasos de un experimento binomial con ocho ensayos.

En un experimento binomial lo que interesa es el *número de éxitos en* n *ensayos*. Si x denota el número de éxitos en n ensayos, es claro que x tomará los valores 0, 1, 2, 3, ..., n. Dado que el número de estos valores es finito, x es una variable aleatoria *discreta*. A la distribución de probabilidad correspondiente a esta variable aleatoria se le llama **distribución de probabilidad binomial**. Por ejemplo, considere el experimento que consiste en lanzar una moneda cinco veces y observar si la cara de la moneda que cae hacia arriba es cara o cruz. Suponga que se desea contar el número de caras que aparecen en los cinco lanzamientos. ¿Presenta este experimento las propiedades de un experimento binomial? ¿Cuál es la variable aleatoria que interesa? Observe que:

1. El experimento consiste en cinco ensayos idénticos; cada ensayo consiste en lanzar una moneda.
2. En cada ensayo hay dos resultados posibles: cara o cruz. Se puede considerar cara como éxito y cruz como fracaso.
3. La probabilidad de éxito y la probabilidad de fracaso son iguales en todos los ensayos, siendo $p = 0.5$ y $1 - p = 0.5$.
4. Los ensayos o lanzamientos son independientes porque al resultado de un ensayo no afecta a lo que pase en los otros ensayos o lanzamientos.

FIGURA 5.2 UNA POSIBLE SUCESIÓN DE ÉXITOS Y FRACASOS EN UN EXPERIMENTO BINOMIAL DE OCHO ENSAYOS

Propiedad 1: El experimento consiste en $n = 8$ ensayos idénticos.

Propiedad 2: En cada ensayo se obtiene como resultado un éxito o un fracaso.

Ensayos →	1	2	3	4	5	6	7	8
Resultados →	S	F	F	S	S	F	S	S

Por tanto, se satisfacen las propiedades de un experimento binomial. La variable aleatoria que interesa es x = número de caras que aparecen en cinco ensayos. En este caso, x puede tomar los valores 0, 1, 2, 3, 4 o 5.

Otro ejemplo, considere a un vendedor de seguros que visita a 10 familias elegidas en forma aleatoria. El resultado correspondiente de la visita a cada familia se clasifica como éxito si la familia compra un seguro y como fracaso si la familia no compra ningún seguro. Por experiencia, el vendedor sabe que la probabilidad de que una familia tomada aleatoriamente compre un seguro es 0.10. Al revisar las propiedades de un experimento binomial aparece que:

1. El experimento consiste en 10 ensayos idénticos; cada ensayo consiste en visitar a una familia.
2. En cada ensayo hay dos resultados posibles: la familia compra un seguro (éxito) o la familia no compra ningún seguro (fracaso).
3. Las probabilidades de que haya compra y de que no haya compra se supone que son iguales en todas las visitas, siendo $p = 0.10$ y $1 - p = 0.90$.
4. Los ensayos son independientes porque las familias se eligen en forma aleatoria.

Como estos cuatro puntos se satisfacen, este ejemplo es un experimento binomial. La variable aleatoria que interesa es el número de ventas al visitar a las 10 familias. En este caso los valores que puede tomar x son 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 y 10.

La propiedad 3 de un experimento binomial se llama *suposición de estacionaridad* y algunas veces se confunde con la propiedad 4, independencia de los ensayos. Para ver la diferencia entre estas dos propiedades, reconsidere el caso del vendedor que visita a las familias para venderles un seguro. Si a medida que el día avanza, el vendedor se va cansando y va perdiendo entusiasmo, la probabilidad de éxito puede disminuir, por ejemplo, a 0.05 en la décima llamada. En tal caso la propiedad 3 (estacionaridad) no se satisface, y no se tiene un experimento binomial. Incluso si la propiedad 4 se satisface —en cada familia la decisión de comprar o no se hizo de manera independiente— si no se satisface la propiedad 3, no se trata de un experimento binomial.

En las aplicaciones de los experimentos binomiales se emplea una fórmula matemática llamada *función de probabilidad binomial* que sirve para calcular la probabilidad de x éxitos en n ensayos. Empleando los conceptos de probabilidad presentados en el capítulo 4, se mostrará, en el contexto de un ilustrativo problema, cómo se desarrolla la fórmula.

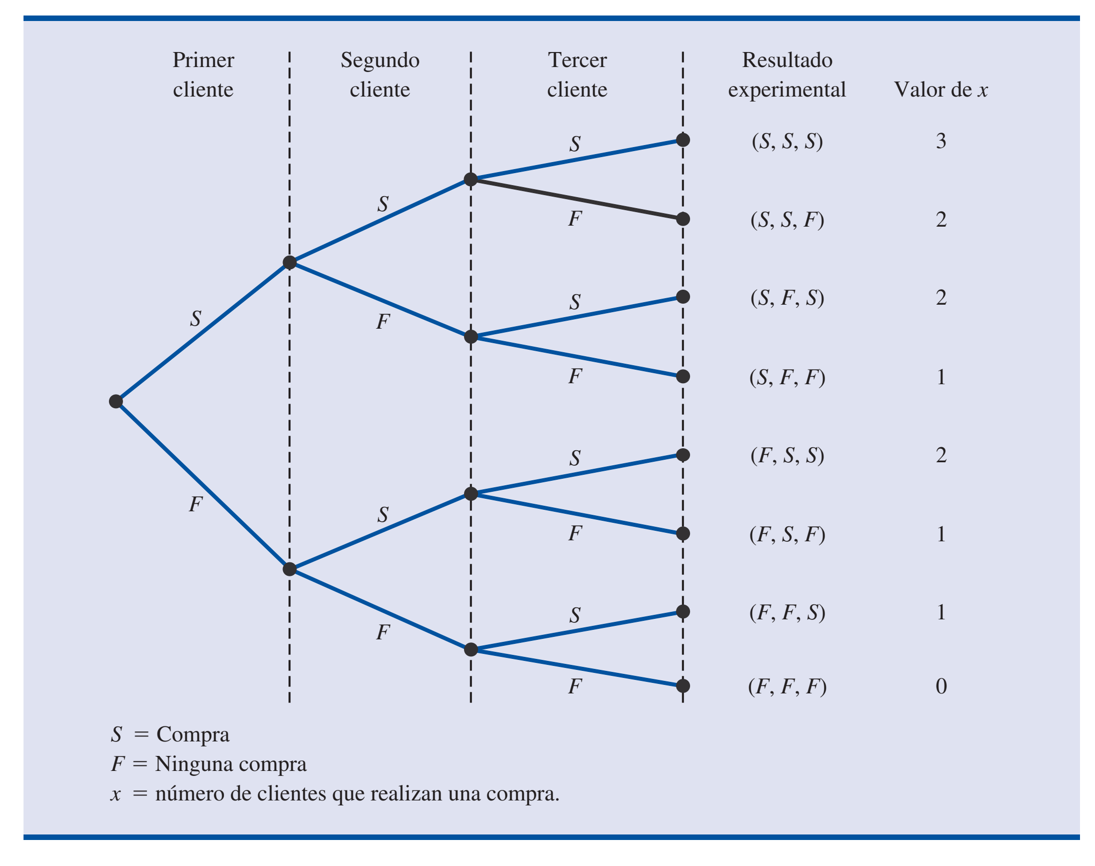
El problema de la tienda de ropa Martin Clothing Store

Considere las decisiones de compra de los próximos tres clientes que lleguen a la tienda de ropa Martin Clothing Store. De acuerdo con la experiencia, el gerente de la tienda estima que la probabilidad de que un cliente realice una compra es 0.30. ¿Cuál es la probabilidad de que dos de los próximos tres clientes realicen una compra?

Un diagrama de árbol (figura 5.3), permite advertir que el experimento de observar a los tres clientes para ver si cada uno de ellos decide realizar una compra tiene ocho posibles resultados. Entonces, si S denota éxito (una compra) y F fracaso (ninguna compra), lo que interesa son los resultados experimentales en los que haya dos éxitos (decisiones de compra) en los tres ensayos. A continuación verifique que el experimento de las tres decisiones de compra es un experimento binomial. Al verificar los cuatro requerimientos de un experimento binomial, se observa que:

1. Es posible describir el experimento como una serie de tres ensayos idénticos, un ensayo por cada uno de los tres clientes que llegan a la tienda.
2. Cada ensayo tiene dos posibles resultados: el cliente hace una compra (éxito) o el cliente no hace ninguna compra (fracaso).
3. La probabilidad de que el cliente haga una compra (0.30) o de que no haga una compra (0.70) se supone que es la misma para todos los clientes.
4. La decisión de comprar de cada cliente es independiente de la decisión de comprar de los otros clientes.

FIGURA 5.3 DIAGRAMA DE ÁRBOL PARA EL PROBLEMA DE LA TIENDA DE ROPA MARTIN CLOTHING STORE



En consecuencia, se satisfacen las propiedades de un experimento binomial.

Con la fórmula siguiente* se calcula el número de resultados experimentales en los que hay exactamente x éxitos en n ensayos.

NÚMERO DE RESULTADOS EXPERIMENTALES EN LOS QUE HAY EXACTAMENTE x ÉXITOS EN n ENSAYOS

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{5.6}$$

donde

$$n! = n(n-1)(n-2)\cdots(2)(1)$$

y por definición,

$$0! = 1$$

Ahora regrese al experimento de las decisiones de compra de tres clientes de la tienda Martin Clothing Store. La ecuación (5.6) sirve para determinar el número de resultados experimentales

* Esta fórmula presentada en el capítulo 4, determina el número de combinaciones de n objetos tomados de x a la vez. En el experimento binomial esta fórmula combinatoria da el número de resultados experimentales (series de n ensayos) en los que hay x éxitos.

en los que hay dos compras; el número de maneras en que son posibles $x = 2$ éxitos en $n = 3$ ensayos. De acuerdo con la ecuación (5.6)

$$\binom{n}{x} = \binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{(3)(2)(1)}{(2)(1)(1)} = \frac{6}{2} = 3$$

La ecuación (5.6) indica que en tres de los resultados experimentales hay dos éxitos. En la figura 5.3 aparecen denotados por (*S, S, F*), (*S, F, S*) y (*F, S, S*).

Empleando la ecuación (5.6) para determinar en cuántos resultados experimentales hay tres éxitos (compras) en tres ensayos, se obtiene

$$\binom{n}{x} = \binom{3}{3} = \frac{3!}{3!(3-3)!} = \frac{3!}{3!0!} = \frac{(3)(2)(1)}{3(2)(1)(1)} = \frac{6}{6} = 1$$

El único resultado experimental con tres éxitos es el identificado por (*S, S, S*) mostrado en la figura 5.3.

Ya sabe que usando la ecuación (5.6) es posible determinar el número de resultados experimentales en los que hay x éxitos. Sin embargo, si va a determinar la probabilidad de x éxitos en n ensayos, es necesario conocer también la probabilidad correspondiente a cada uno de estos resultados experimentales. Como en un experimento binomial, los ensayos son independientes, para hallar la probabilidad de una determinada sucesión de éxitos y fracasos simplemente se multiplican las probabilidades correspondientes al resultado de cada ensayo.

La probabilidad de que los dos primeros clientes compren y el tercero no compre, denotada por (*S, S, F*) está dada por

$$pp(1 - p)$$

Puesto que la probabilidad de compra en cualquier ensayo es 0.30, la probabilidad de que haya una compra en los dos primeros ensayos y que no haya compra en el tercer ensayo es

$$(0.30)(0.30)(0.70) = (0.30)^2(0.70) = 0.063$$

Hay otros dos resultados experimentales en los que también se obtienen dos éxitos y un fracaso. A continuación se presentan las probabilidades de los tres resultados experimentales en los que hay dos éxitos.

Resultados de los ensayos				
1er. cliente	**2o. cliente**	**3er. cliente**	**Resultado experimental**	**Probabilidad de este resultado experimental**
Compra	Compra	No hay compra	(S, S, F)	$pp(1-p) = p^2(1-p)$ $= (0.30)^2(0.70) = 0.063$
Compra	Compra	Compra	(S, F, S)	$p(1-p)p = p^2(1-p)$ $= (0.30)^2(0.70) = 0.063$
No hay compra	Compra	Compra	(F, S, S)	$(1-p)pp = p^2(1-p)$ $= (0.30)^2(0.70) = 0.063$

Observe que los tres resultados experimentales en los que hay dos éxitos tienen la misma probabilidad. Esto se cumple en general. En cualquier experimento binomial todas las series de resultados de ensayos en las que hay x éxitos en n ensayos tienen la *misma probabilidad* de ocurrencia. A continuación se presenta la probabilidad de cada una de las series de ensayos en las que hay x éxitos en n ensayos.

$$\begin{array}{l}\text{Probabilidad de una} \\ \text{determinada serie de} \\ \text{resultados de ensayos}\end{array} = p^x(1-p)^{(n-x)}$$

En el caso de la tienda de ropa Martin Clothing Store, esta fórmula indica que la probabilidad de cualquier resultado experimental con dos éxitos es $p^2(1-p)^{(3-2)} = p^2(1-p)^1 = (0.30)^2(0.70^1) = 0.63$.

Como la ecuación (5.6) da el número de resultados de un experimento binomial en el que hay x éxitos, y la ecuación (5.7) da la probabilidad de cada serie en la que hay x éxitos, combinando las ecuaciones (5.6) y (5.7) se obtiene la **función de probabilidad binomial** siguiente.

FUNCIÓN DE PROBABILIDAD BINOMIAL

$$f(x) = \binom{n}{x} p^x(1-p)^{(n-x)} \qquad \textbf{(5.8)}$$

donde

$$\begin{aligned} f(x) &= \text{probabilidad de } x \text{ éxitos en } n \text{ ensayos} \\ n &= \text{número de ensayos} \\ \binom{n}{x} &= \frac{n!}{x!(n-x)!} \\ p &= \text{probabilidad de un éxito en cualquiera de los ensayos} \\ 1-p &= \text{probabilidad de un fracaso en cualquiera de los ensayos} \end{aligned}$$

En el ejemplo de la tienda de ropa Martin Clothing Store se calculará ahora la probabilidad de que ningún cliente realice una compra, de que exactamente un cliente realice una compra, de que exactamente dos clientes realicen una compra y de que los tres clientes realicen una compra. Los cálculos se presentan en forma resumida en la tabla 5.7, que da la distribución de probabilidad para el número de clientes que hacen una compra. La figura 5.4 es una gráfica de esta distribución de probabilidad.

La función de probabilidad binomial es aplicable a *cualquier* experimento binomial. Si encuentra que una situación presenta las propiedades de un experimento binomial y conoce los valores de n y p, use la ecuación (5.8) para calcular la probabilidad de x éxitos en n ensayos.

TABLA 5.7 DISTRIBUCIÓN DE PROBABILIDAD BINOMIAL PARA EL NÚMERO DE CLIENTES QUE HACEN UNA COMPRA

x	$f(x)$
0	$\frac{3!}{0!3!}(0.30)^0(0.70)^3 = 0.343$
1	$\frac{3!}{1!2!}(0.30)^1(0.70)^2 = 0.441$
2	$\frac{3!}{2!1!}(0.30)^2(0.70)^1 = 0.189$
3	$\frac{3!}{3!0!}(0.30)^3(0.70)^0 = 0.027$
	1.000

FIGURA 5.4 REPRESENTACIÓN GRÁFICA DE LA DISTRIBUCIÓN DE PROBABILIDAD BINOMIAL PARA EL NÚMERO DE CLIENTES QUE HACEN UNA COMPRA

Si considera variaciones del experimento de la tienda de ropa, por ejemplo, que lleguen a la tienda 10 clientes en lugar de tres clientes, también se emplea la función de probabilidad binomial dada por la ecuación (5.8). Suponga que tiene un experimento binomial con $n = 10$, $x = 4$ y $p = 0.30$. La probabilidad de que cuatro de los 10 clientes que entran en la tienda de ropa realicen una compra es

$$f(4) = \frac{10!}{4!6!}(0.30)^4(0.70)^6 = 0.2001$$

Uso de las tablas de probabilidades binomiales

Existen tablas que dan la probabilidad de x éxitos en n ensayos de un experimento binomial. Estas tablas son fáciles de usar y los resultados se obtienen más rápidamente que con la ecuación (5.8). La tabla 5 del apéndice B es una de estas tablas de probabilidades binomiales. Una parte de esta tabla se presenta en la tabla 5.8. Para usarla es necesario especificar los valores de n, p y x en el experimento binomial de que se trate. En el ejemplo que se presenta en la parte superior de la tabla 5.8 se ve que la probabilidad de $x = 3$ éxitos en un experimento binomial con $n = 10$ y $p = 0.40$ es 0.2150. Use la ecuación (5.8) para verificar que este mismo resultado se obtiene si usa la función de probabilidad binomial directamente.

Ahora se usará la tabla 5.8 para corroborar la probabilidad de 4 éxitos en 10 ensayos en el problema de la tienda de ropa Martin Clothing Store. Observe que el valor de $f(4) = 0.2001$ se lee directamente de la tabla de probabilidades binomiales, eligiendo $n = 10$, $x = 4$ y $p = 0.30$.

Con las calculadoras modernas estas tablas son casi innecesarias. Es muy fácil evaluar la ecuación (5.8) directamente.

Aun cuando las tablas de probabilidades binomiales son relativamente fáciles de utilizar, es imposible contar con tablas que tengan todos los valores de n y p de un experimento binomial. Sin embargo, con las calculadoras de hoy en día, usar la ecuación (5.8) para calcular la probabilidad deseada no es difícil, en especial si el número de ensayos no es grande. En los ejercicios tendrá la oportunidad de usar la ecuación (5.8) para calcular probabilidades binomiales, a menos que el problema pida que use la tabla de probabilidad binomial.

TABLA 5.8 ALGUNOS VALORES DE LA TABLA DE PROBABILIDAD BINOMIAL
EJEMPLO: $n = 10, x = 3, p = 0.40; f(3) = 0.2150$

		p									
n	***x***	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
9	0	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.2985	0.3874	0.3679	0.3020	0.2253	0.1556	0.1004	0.0605	0.0339	0.0176
	2	0.0629	0.1722	0.2597	0.3020	0.3003	0.2668	0.2162	0.1612	0.1110	0.0703
	3	0.0077	0.0446	0.1069	0.1762	0.2336	0.2668	0.2716	0.2508	0.2119	0.1641
	4	0.0006	0.0074	0.0283	0.0661	0.1168	0.1715	0.2194	0.2508	0.2600	0.2461
	5	0.0000	0.0008	0.0050	0.0165	0.0389	0.0735	0.1181	0.1672	0.2128	0.2461
	6	0.0000	0.0001	0.0006	0.0028	0.0087	0.0210	0.0424	0.0743	0.1160	0.1641
	7	0.0000	0.0000	0.0000	0.0003	0.0012	0.0039	0.0098	0.0212	0.0407	0.0703
	8	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0013	0.0035	0.0083	0.0176
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0020
10	0	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1	0.3151	0.3874	0.3474	0.2684	0.1877	0.1211	0.0725	0.0403	0.0207	0.0098
	2	0.0746	0.1937	0.2759	0.3020	0.2816	0.2335	0.1757	0.1209	0.0763	0.0439
	3	0.0105	0.0574	0.1298	0.2013	0.2503	0.2668	0.2522	0.**2150**	0.1665	0.1172
	4	0.0010	0.0112	0.0401	0.0881	0.1460	0.2001	0.2377	0.2508	0.2384	0.2051
	5	0.0001	0.0015	0.0085	0.0264	0.0584	0.1029	0.1536	0.2007	0.2340	0.2461
	6	0.0000	0.0001	0.0012	0.0055	0.0162	0.0368	0.0689	0.1115	0.1596	0.2051
	7	0.0000	0.0000	0.0001	0.0008	0.0031	0.0090	0.0212	0.0425	0.0746	0.1172
	8	0.0000	0.0000	0.0000	0.0001	0.0004	0.0014	0.0043	0.0106	0.0229	0.0439
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0016	0.0042	0.0098
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010

Los paquetes de software para estadística como Minitab y los paquetes de hojas de cálculo como Excel también están habilitadas para calcular probabilidades binomiales. Considere el ejemplo de la tienda de ropa Martin Clothing Store con $n = 10$ y $p = 0.30$. En la figura 5.5 se muestran las probabilidades binomiales para todos los valores posibles de x, generadas por Minitab. Observe que estos valores son los mismos que se encuentran en la columna $p = 0.30$ de la tabla 5.8. En el apéndice 5.1 se da paso por paso el procedimiento en Minitab para generar el resultado que se muestra en la figura 5.5. En el apéndice 5.2 se describe cómo usar Excel para calcular probabilidades binomiales.

Valor esperado y varianza en la distribución binomial

En la sección 5.3 se dieron las fórmulas para calcular el valor esperado y la varianza de una variable aleatoria discreta. En el caso especial de que la variable aleatoria tenga una distribución binomial para la que se conoce el número de ensayos n y la probabilidad de éxito p, las fórmulas generales para el valor esperado y la varianza se simplifican. El resultado se muestra a continuación.

VALOR ESPERADO Y VARIANZA EN LA DISTRIBUCIÓN BINOMIAL

$$E(x) = \mu = np \quad \textbf{(5.9)}$$

$$\text{Var}(x) = \sigma^2 = np(1 - p) \quad \textbf{(5.10)}$$

FIGURA 5.5 RESULTADOS DE MINITAB QUE MUESTRAN LAS PROBABILIDADES BINOMIALES PARA EL PROBLEMA DE LA TIENDA DE ROPA MARTIN CLOTHING STORE

x	P(X = x)
0.00	0.0282
1.00	0.1211
2.00	0.2335
3.00	0.2668
4.00	0.2001
5.00	0.1029
6.00	0.0368
7.00	0.0090
8.00	0.0014
9.00	0.0001
10.00	0.0000

Para el problema de los tres clientes de la tienda de ropa Martin Clothing Store, use la ecuación (5.9) para calcular el número esperado de clientes que harán una compra.

$$E(x) = np = 3(0.30) = 0.9$$

Suponga que Martin Clothing Store pronostica que el mes próximo 1000 clientes visitarán la tienda. ¿Cuál es el número esperado de clientes que harán una compra? La respuesta es $\mu = np = (1000)(0.30) = 300$. Así, para aumentar el número esperado de compras, Martin debe hacer que más clientes visiten su tienda o de alguna manera aumentar la probabilidad de que una persona que visite la tienda haga una compra.

En el caso de los tres clientes de la tienda de ropa Martin Clothing Store, la varianza y la desviación estándar del número de clientes que harán una compra son

$$\sigma^2 = np(1 - p) = 3(0.3)(0.7) = 0.63$$
$$\sigma = \sqrt{0.63} = 0.79$$

Para los próximos 1000 clientes que visiten la tienda, la varianza y la desviación estándar del número de clientes que harán una compra son

$$\sigma^2 = np(1 - p) = 1000(0.3)(0.7) = 210$$
$$\sigma = \sqrt{210} = 14.49$$

NOTAS Y COMENTARIOS

1. En las tablas binomiales del apéndice B los valores de p llegan sólo hasta 0.50. Es posible pensar que estas tablas no son útiles cuando la probabilidad de éxito es mayor a 0.50. Sin embargo, puede usarlas observando que la probabilidad de $n - x$ fracasos es también la probabilidad de x éxitos. Cuando la probabilidad de éxito es mayor que $p = 0.50$, en lugar de la probabilidad de éxito calcule la probabilidad de $n - x$ fracasos. Cuando $p > 0.50$, la probabilidad de fracaso, $1 - p$, será menor que 0.50.

2. En algunas fuentes se presentan tablas binomiales en forma acumulada. Al usar estas tablas para hallar la probabilidad de x éxitos en n ensayos hay que hacer una resta. Por ejemplo, $f(2) = P(x \leq 2) - P(x \leq 1)$. Las tablas que se presentan en este libro dan estas probabilidades. Para calcular probabilidades acumuladas usando las tablas de este libro, sume las probabilidades individuales. Por ejemplo, para calcular $P(x \leq 2)$ usando las tablas del libro, sume $f(0) + f(1) + f(2)$.

Ejercicios

Métodos



25. Considere un experimento binomial con dos ensayos y $p = 0.4$.
    a. Dibuje un diagrama de árbol para este experimento (véase figura 5.3).
    b. Calcule la probabilidad de un éxito, $f(1)$.
    c. Calcule $f(0)$.
    d. Calcule $f(2)$.
    e. Calcule la probabilidad de por lo menos un éxito.
    f. Calcule el valor esperado, la varianza y la desviación estándar.

26. Considere un experimento binomial con $n = 10$ y $p = 0.10$.
    a. Calcule $f(0)$.
    b. Calcule $f(2)$.
    c. Calcule $P(x \leq 2)$.
    d. Calcule $P(x \geq 1)$.
    e. Calcule $E(x)$.
    f. Calcule $\text{Var}(x)$ y σ.

27. Considere un experimento binomial con $n = 20$ y $p = 0.70$.
    a. Calcule $f(12)$.
    b. Calcule $f(16)$.
    c. Calcule $P(x \geq 16)$.
    d. Calcule $P(x \leq 15)$.
    e. Calcule $E(x)$.
    f. Calcule $\text{Var}(x)$ y σ.

Aplicaciones

28. Una encuesta de Harris Interactive para InterContinental Hoteld and Resorts preguntó: "Cuando viaja al extranjero, ¿suele aventurarse usted solo para conocer la cultura o prefiere permanecer con el grupo de su *tour* y apegarse al itinerario?" Se encontró que 23% prefiere permanecer con el grupo de su *tour* (*USA Today,* 21 de enero de 2004).
    a. ¿Cuál es la probabilidad de que en una muestra de seis viajeros, dos prefieran permanecer con su grupo?
    b. ¿De que en una muestra de seis viajeros, por lo menos dos prefieran permanecer con su grupo?
    c. ¿De que en una muestra de 10 viajeros, ninguno prefiera permanecer con su grupo?

29. En San Francisco, 30% de los trabajadores emplean el transporte público (*USA Today,* 21 de diciembre de 2005).
    a. ¿Cuál es la probabilidad de que en una muestra de 10 trabajadores exactamente tres empleen el transporte público?
    b. ¿De que en una muestra de 10 trabajadores por lo menos tres empleen el transporte público?



30. Cuando una máquina nueva funciona adecuadamente, sólo 3% de los artículos producidos presentan algún defecto. Suponga que selecciona aleatoriamente dos piezas producidas con la nueva máquina y que busca el número de piezas defectuosas.
    a. Describa las condiciones en las que éste será un experimento binomial.
    b. Elabore un diagrama de árbol como el de la figura 5.3 en el que se muestre este problema como un experimento de dos ensayos.
    c. ¿En cuántos resultados experimentales hay exactamente una pieza defectuosa?
    d. Calcule las probabilidades de hallar ninguna pieza defectuosa, exactamente una pieza defectuosa y dos piezas defectuosas.

31. Nueve por ciento de los estudiantes tienen un balance en su tarjeta de crédito mayor a $7000 (*Reader's Digest*, julio de 2002). Suponga que selecciona aleatoriamente 10 estudiantes para entrevistarlos respecto del uso de su tarjeta de crédito

a. ¿Es la selección de 10 estudiantes un experimento binomial? Explique.
b. ¿Cuál es la probabilidad de que dos de los estudiantes tengan un balance en su tarjeta de crédito superior a \$7000?
c. ¿De que ninguno tenga un balance en su tarjeta de crédito superior a \$7000?
d. ¿De que por lo menos tres tengan un balance en su tarjeta de crédito superior a \$7000?

32. Los radares militares y los sistemas para detección de misiles tienen por objeto advertir a un país de un ataque enemigo. Una cuestión de confiabilidad es si el sistema de detección será capaz de detectar un ataque y emitir un aviso. Suponga que la probabilidad de que un determinado sistema de detección detecte un ataque con misiles es 0.90. Use la distribución de probabilidad binomial para responder las preguntas siguientes.
 a. ¿Cuál es la probabilidad de que un solo sistema de detección detecte un ataque?
 b. Si se instalan dos sistemas de detección en una misma área y los dos operan independientemente, ¿cuál es la probabilidad de que por lo menos uno de los sistemas detecte el ataque?
 c. Si se instalan tres sistemas, ¿cuál es la probabilidad de que por lo menos uno de los sistemas detecte el ataque?
 d. ¿Recomendaría que se usaran varios sistemas de detección? Explique.
33. Cincuenta por ciento de los estadounidenses creyeron que el país se encontraba en una recesión aun cuando en la economía no se habían observado dos trimestres seguidos con crecimiento negativo. (*BusinessWeek,* 30 de julio de 2001). Dada una muestra de 20 estadounidenses, calcule lo siguiente.
 a. Calcule la probabilidad de que exactamente 12 personas hayan creído que el país estaba en recesión.
 b. De que no más de cinco personas hayan creído que el país estaba en recesión
 c. ¿Cuántas personas esperaría usted que dijeran que el país estuvo en recesión?
 d. Calcule la varianza y la desviación estándar del número de personas que creyeron que el país estuvo en recesión.
34. En una encuesta realizada por la Oficina de Censos de Estados Unidos se encontró que 25% de las personas de 25 años o más habían estudiado cuatro años en la universidad (*The New York Times Almanac,* 2006). Dada una muestra de 15 individuos de 25 años o más, conteste las preguntas siguientes.
 a. ¿Cuál es la probabilidad de que cuatro hayan estudiado cuatro años en la universidad?
 b. ¿De que tres o más hayan estudiado cuatro años en la universidad?
35. En una universidad se encontró que 20% de los estudiantes no terminan el primer curso de estadística, al curso se inscriben 20 estudiantes.
 a. Calcule la probabilidad de que dos o menos no terminen.
 b. De que cuatro, exactamente, no terminen.
 c. De que más de tres no terminen.
 d. ¿Cuál es el número esperado de estudiantes que no terminan?
36. En el caso particular de una variable aleatoria binomial, es factible calcular la varianza empleando la fórmula $\sigma^2 = np(1 - p)$. En el caso del problema de la tienda de ropa Martin Clothing Store, en donde $n = 3$ y $p = 0.3$, se encontró que $\sigma^2 = np(1 - p) = 3(0.3)(0.7) = 0.63$. Aplique la definición general de varianza para una variable aleatoria discreta, ecuación (5.5), y las probabilidades de la tabla 5.7 para comprobar que la varianza es 0.63
37. Veintitrés por ciento de los automóviles no cuenta con un seguro (CNN, 23 de febrero de 2006). En un fin de semana determinado hay 35 automóviles que sufren un accidente.
 a. ¿Cuál es el número esperado de estos automóviles que no cuentan con un seguro?
 b. ¿Cuál es la varianza y la desviación estándar?

5.5 Distribución de probabilidad de Poisson

En esta sección estudiará una variable aleatoria discreta que se suele usar para estimar el número de veces que sucede un hecho determinado (ocurrencias) en un intervalo de tiempo o de espacio. Por ejemplo, la variable de interés va desde el número de automóviles que llegan (llegadas) a un lavado de coches en una hora o el número de reparaciones necesarias en 10 millas de una autopista hasta el número de fugas en 100 millas de tubería. Si se satisfacen las condiciones si-

La distribución de probabilidad de Poisson suele emplearse para modelar las llegadas aleatorias a una línea de espera (fila).

guientes, el número de ocurrencias es una variable aleatoria discreta, descrita por la **distribución de probabilidad de Poisson**.

PROPIEDADES DE UN EXPERIMENTO DE POISSON

1. La probabilidad de ocurrencia es la misma para cualesquiera dos intervalos de la misma magnitud.
2. La ocurrencia o no-ocurrencia en cualquier intervalo es independiente de la ocurrencia o no-ocurrencia en cualquier otro intervalo.

La **función de probabilidad de Poisson** se define mediante la ecuación (5.11).

Simeon Poisson dio clases de matemáticas en la Ecole Polytechnique de París de 1802 a 1808. En 1837 publicó un trabajo titulado "Investigación sobre la probabilidad de veredictos en materia criminal y civil" en el que presenta un estudio sobre lo que después se conoció como distribución de Poisson.

FUNCIÓN DE PROBABILIDAD DE POISSON

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \tag{5.11}$$

en donde

$f(x)$ = probabilidad de x ocurrencias en un intervalo
μ = valor esperado o número medio de ocurrencias en un intervalo
e = 2.71828

Antes de considerar un ejemplo para ver cómo se usa la distribución de Poisson, observe que el número de ocurrencias x, no tiene límite superior. Ésta es una variable aleatoria discreta que toma los valores de una sucesión infinita de números ($x = 0, 1, 2, \ldots$).

Un ejemplo considerando intervalos de tiempo

Los laboratorios Bell usaron la distribución de Poisson para modelar las llegadas de llamadas telefónicas.

Suponga que desea saber el número de llegadas, en un lapso de 15 minutos, a la rampa del cajero automático de un banco. Si se puede suponer que la probabilidad de llegada de los automóviles es la misma en cualesquiera dos lapsos de la misma duración y si la llegada o no–llegada de un automóvil en cualquier lapso es independiente de la llegada o no–llegada de un automóvil en cualquier otro lapso, se puede aplicar la función de probabilidad de Poisson. Dichas condiciones se satisfacen y en un análisis de datos pasados encuentra que el número promedio de automóviles que llegan en un lapso de 15 minutos es 10; en este caso use la función de probabilidad siguiente.

$$f(x) = \frac{10^x e^{-10}}{x!}$$

Aquí la variable aleatoria es x = número de automóviles que llegan en un lapso de 15 minutos.

Si la administración desea saber la probabilidad de que lleguen exactamente cinco automóviles en 15 minutos, $x = 5$, y se obtiene

$$\text{Probabilidad de que lleguen exactamente 5 automóviles en 15 minutos} = f(5) = \frac{10^5 e^{-10}}{5!} = 0.0378$$

Aunque esta probabilidad se obtuvo evaluando la función de probabilidad con $\mu = 10$ y $x = 5$, suele ser más fácil consultar una tabla de probabilidad de Poisson. Dichas tablas proporcionan las probabilidades para valores específicos de x y μ. La tabla 7 del apéndice B es una tabla de probabilidad de Poisson. Para mayor comodidad, en la tabla 5.9 se reproduce parte de la tabla 7 del apéndice B. Observe que para usar una tabla de probabilidades de Poisson se necesitan sólo

TABLA 5.9 ALGUNOS VALORES DE LAS TABLAS DE PROBABILIDAD DE POISSON
EJEMPLO: $\mu = 10, x = 5; f(5) = .0378$

	μ									
x	**9.1**	**9.2**	**9.3**	**9.4**	**9.5**	**9.6**	**9.7**	**9.8**	**9.9**	**10**
0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
1	0.0010	0.0009	0.0009	0.0008	0.0007	0.0007	0.0006	0.0005	0.0005	0.0005
2	0.0046	0.0043	0.0040	0.0037	0.0034	0.0031	0.0029	0.0027	0.0025	0.0023
3	0.0140	0.0131	0.0123	0.0115	0.0107	0.0100	0.0093	0.0087	0.0081	0.0076
4	0.0319	0.0302	0.0285	0.0269	0.0254	0.0240	0.0226	0.0213	0.0201	0.0189
5	0.0581	0.0555	0.0530	0.0506	0.0483	0.0460	0.0439	0.0418	0.0398	0.**0378**
6	0.0881	0.0851	0.0822	0.0793	0.0764	0.0736	0.0709	0.0682	0.0656	0.0631
7	0.1145	0.1118	0.1091	0.1064	0.1037	0.1010	0.0982	0.0955	0.0928	0.0901
8	0.1302	0.1286	0.1269	0.1251	0.1232	0.1212	0.1191	0.1170	0.1148	0.1126
9	0.1317	0.1315	0.1311	0.1306	0.1300	0.1293	0.1284	0.1274	0.1263	0.1251
10	0.1198	0.1210	0.1219	0.1228	0.1235	0.1241	0.1245	0.1249	0.1250	0.1251
11	0.0991	0.1012	0.1031	0.1049	0.1067	0.1083	0.1098	0.1112	0.1125	0.1137
12	0.0752	0.0776	0.0799	0.0822	0.0844	0.0866	0.0888	0.0908	0.0928	0.0948
13	0.0526	0.0549	0.0572	0.0594	0.0617	0.0640	0.0662	0.0685	0.0707	0.0729
14	0.0342	0.0361	0.0380	0.0399	0.0419	0.0439	0.0459	0.0479	0.0500	0.0521
15	0.0208	0.0221	0.0235	0.0250	0.0265	0.0281	0.0297	0.0313	0.0330	0.0347
16	0.0118	0.0127	0.0137	0.0147	0.0157	0.0168	0.0180	0.0192	0.0204	0.0217
17	0.0063	0.0069	0.0075	0.0081	0.0088	0.0095	0.0103	0.0111	0.0119	0.0128
18	0.0032	0.0035	0.0039	0.0042	0.0046	0.0051	0.0055	0.0060	0.0065	0.0071
19	0.0015	0.0017	0.0019	0.0021	0.0023	0.0026	0.0028	0.0031	0.0034	0.0037
20	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012	0.0014	0.0015	0.0017	0.0019
21	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0007	0.0008	0.0009
22	0.0001	0.0001	0.0002	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004
23	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001

dos valores, x y μ. En la tabla 5.9 la probabilidad de cinco llegadas en un lapso de 15 minutos se obtiene localizando el valor que se encuentra en el renglón correspondiente a $x = 5$ y la columna correspondiente a $\mu = 10$. Así obtiene $f(5) = 0.0378$

Una propiedad de la distribución de Poisson es que la media y la varianza son iguales.

La media de la distribución de Poisson en el ejemplo anterior fue $\mu = 10$ llegadas en un lapso de 15 minutos. Una propiedad de la distribución de Poisson es que la media y la varianza de la distribución son *iguales*. Por tanto, la varianza del número de llegadas en un lapso de 15 minutos es $\sigma^2 = 10$. La desviación estándar es $\sigma = \sqrt{10} = 3.16$.

En el ejemplo anterior se usó un lapso de 15 minutos, pero también se usan otros lapsos. Suponga que desea calcular la probabilidad de una llegada en un lapso de 3 minutos. Como 10 es el número esperado de llegadas en un lapso de 15 minutos: $10/15 = 2/3$ es el número esperado de llegadas en un lapso de un minuto y que $(2/3)(3 \text{ minutos}) = 2$ es el número esperado de llegadas en un lapso de 3 minutos. Entonces, la probabilidad de x llegadas en un lapso de 3 minutos con $\mu = 2$ está dada por la siguiente función de probabilidad de Poisson.

$$f(x) = \frac{2^x e^{-2}}{x!}$$

La probabilidad de una llegada en un lapso de 3 minutos se obtiene como sigue:

$$\text{Probabilidad de exactamente una llegada en 3 minutos} = f(1) = \frac{2^1 e^{-2}}{1!} = 0.2707$$

Antes se calculó la probabilidad de cinco llegadas en un lapso de 15 minutos; se obtuvo 0.0378. Observe que la probabilidad de una llegada en un lapso de tres minutos (0.2707) no es la misma. Para calcular la probabilidad de Poisson en un lapso diferente, primero hay que convertir la llegada media al lapso que interesa y después calcular la probabilidad.

Un ejemplo considerando intervalos de longitud o de distancia

Ahora se da un ejemplo en el que no aparecen intervalos de tiempo y en el que se usa la distribución de Poisson. Asuma que le interesa la ocurrencia de una avería importante en una autopista un mes después de que ha sido repavimentada. Supondrá que la probabilidad de que haya una avería es la misma en cualesquiera dos tramos, de una misma longitud, de la autopista y que la ocurrencia o no–ocurrencia de una avería en un tramo es independiente de la ocurrencia o no-ocurrencia de una avería en cualquier otro tramo. Por tanto, emplea la distribución de Poisson.

También sabe que el promedio de averías importantes, un mes después de la repavimentación, son dos averías por milla. Desea determinar la probabilidad de que no haya ninguna avería en un determinado tramo de tres millas de autopista. Como lo que interesa es un intervalo cuya longitud es de tres millas, $\mu = (2 \text{ averías/milla})(3 \text{ millas}) = 6$ representa el número esperado de averías importantes en un tramo de tres millas de autopista. Mediante la ecuación (5.11), la probabilidad de que no haya ninguna avería importante es $f(0) = 6^0 e^{-6}/0! = 0.0025$. Por tanto, es poco probable que no haya ninguna avería importante en este tramo de tres millas. En efecto, este ejemplo indica que hay una probabilidad de $1 - 0.0025 = 0.9975$ de que haya por lo menos una avería importante en este tramo de tres millas de autopista.

Ejercicios

Métodos

38. Considere una distribución de Poisson con $\mu = 3$.
    a. Dé la adecuada función de probabilidad de Poisson.
    b. Calcule $f(2)$.
    c. Calcule $f(1)$.
    d. Calcule $P(x \geq 2)$.



39. Considere una distribución de Poisson en que la media es de dos ocurrencias por un periodo de tiempo.
    a. Dé la adecuada función de probabilidad de Poisson.
    b. ¿Cuál es el número esperado de ocurrencias en tres periodos de tiempo?
    c. Dé la adecuada función de probabilidad de Poisson para determinar la probabilidad de x ocurrencias en tres lapsos.
    d. Calcule la probabilidad de dos ocurrencias en un periodo de tiempo.
    e. Calcule la probabilidad de seis ocurrencias en tres periodos de tiempo.
    f. Calcule la probabilidad de cinco ocurrencias en dos periodos de tiempo.

Aplicaciones

40. A la oficina de reservaciones de una aerolínea regional llegan 48 llamadas por hora.
    a. Calcule la probabilidad de recibir cinco llamadas en un lapso de 5 minutos.
    b. Estime la probabilidad de recibir exactamente 10 llamadas en un lapso de 15 minutos.
    c. Suponga que no hay ninguna llamada en espera. Si el agente de viajes necesitará 5 minutos para la llamada que está atendiendo, ¿cuántas llamadas habrá en espera para cuando él termine? ¿Cuál es la probabilidad de que no haya ninguna llamada en espera?
    d. Si en este momento no hay ninguna llamada, ¿cuál es la probabilidad de que el agente de viajes pueda tomar 3 minutos de descanso sin ser interrumpido por una llamada?

41. Durante el periodo en que una universidad recibe inscripciones por teléfono, llegan llamadas a una velocidad de una cada dos minutos.
    a. ¿Cuál es el número esperado de llamadas en una hora?
    b. ¿Cuál es la probabilidad de que haya tres llamadas en cinco minutos?
    c. ¿De que no haya llamadas en un lapso de cinco minutos?



42. En Estados Unidos, cada año, más de 50 millones de huéspedes se alojan en un "Bread and breakfast" (B&B). El sitio Web dedicado a los alojamientos tipo Bread and Breakfast en Estados Unidos (www.bestinns.net), que tiene un promedio aproximado de siete visitantes por minuto, permite a muchos B&B obtener huéspedes (*Time*, septiembre de 2001).
    a. Calcule la probabilidad de que no haya ningún visitante al sitio Web en un lapso de un minuto.
    b. De que haya dos o más visitantes al sitio Web en un lapso de un minuto.
    c. De que haya uno o más visitantes al sitio Web en un lapso de 30 segundos.
    d. De que haya cinco o más visitantes al sitio Web en un lapso de un minuto.

43. Los pasajeros de las aerolíneas llegan en forma aleatoria e independiente al mostrador de revisión de pasajeros. La tasa media de llegada es 10 pasajeros por minuto.
    a. Calcule la probabilidad de que no llegue ningún pasajero en un lapso de un minuto.
    b. Calcule la probabilidad de que lleguen tres o menos pasajeros en un lapso de un minuto.
    c. De que no llegue ningún pasajero en un lapso de 15 segundos.
    d. De que llegue por lo menos un pasajero en un lapso de 15 segundos.

44. Cada año ocurren en promedio 15 accidentes aéreos (*The World Almanac and Book of Facts*, 2004).
    a. Calcule el número medio de accidentes aéreos por mes.
    b. Calcule la probabilidad de que no haya ningún accidente en un mes.
    c. De que haya exactamente un accidente en un mes.
    d. De que haya más de un accidente en un mes.

45. El National Safety Council de Estados Unidos estima que los accidentes fuera del trabajo tienen para las empresas un costo de casi $200 mil millones anuales en pérdida de productividad. Con base en estos datos, las empresas que tienen 50 empleados esperan tener por lo menos tres accidentes fuera del trabajo por año. Para estas empresas con 50 empleados, conteste las preguntas siguientes.
    a. ¿Cuál es la probabilidad de que no haya ningún accidente fuera del trabajo en un año?
    b. ¿De que haya por lo menos dos accidentes fuera del trabajo en un año?
    c. ¿Cuál es el número esperado de accidentes fuera del trabajo en un lapso de seis meses?
    d. ¿Cuál es la probabilidad de que no haya ningún accidente fuera del trabajo en los próximos seis meses?



Distribución de probabilidad hipergeométrica

La **distribución de probabilidad hipergeométrica** está estrechamente relacionada con la distribución binomial. Pero difieren en dos puntos: en la distribución hipergeométrica los ensayos no son independientes y la probabilidad de éxito varía de ensayo a ensayo.

En la notación usual en la distribución hipergeométrica, r denota el número de elementos considerados como éxitos que hay en una población de tamaño N, y $N - r$ denota el número de elementos considerados como fracasos que hay en dicha población. La **función de probabilidad hipergeométrica** se usa para calcular la probabilidad de que en una muestra aleatoria de n elementos, seleccionados sin reemplazo, se tengan x éxitos y $n - x$ fracasos. Para que se presente este resultado, debe tener x éxitos de los r éxitos que hay en la población y $n - x$ fracasos de los $N - r$ fracasos. La siguiente función de probabilidad hipergeométrica proporciona $f(x)$, la probabilidad de tener x éxitos en una muestra de tamaño n.

FUNCIÓN DE PROBABILIDAD HIPERGEOMÉTRICA

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \qquad \text{para } 0 \leq x \leq r \tag{5.12}$$

donde

$f(x)$ = probabilidad de x éxitos en n ensayos
n = número de ensayos
N = número de elementos en la población
r = número de elementos en la población considerados como éxitos

Observe que $\binom{N}{n}$ representa el número de maneras en que es posible tomar una muestra de tamaño n de una población de tamaño N; $\binom{r}{x}$ representa el número de formas en que se toman x éxitos de un total de r éxitos que hay en la población, y $\binom{N-r}{n-x}$ representa el número de maneras en que se puede tomar $n - x$ fracasos de un total de $N - r$ que hay en la población.

Para ilustrar los cálculos que se emplean al usar la ecuación (5.12), considere la siguiente aplicación al control de calidad. Una empresa fabrica fusibles que empaca en cajas de 12 unidades cada una. Asuma que un inspector selecciona al azar tres de los 12 fusibles de una caja para inspeccionarlos. Si la caja contiene exactamente cinco fusibles defectuosos, ¿cuál es la probabilidad de que el inspector encuentre que uno de los tres fusibles está defectuoso? En esta aplicación $n = 3$ y $N = 12$. Si $r = 5$ fusibles defectuosos en la caja, la probabilidad de hallar $x = 1$ defectuoso es

$$f(1) = \frac{\binom{5}{1}\binom{7}{2}}{\binom{12}{3}} = \frac{\left(\frac{5!}{1!4!}\right)\left(\frac{7!}{2!5!}\right)}{\left(\frac{12!}{3!9!}\right)} = \frac{(5)(21)}{220} = 0.4773$$

Ahora suponga que desea conocer la probabilidad de hallar *por lo menos* un fusible defectuoso. La manera más sencilla de contestar es calcular primero la probabilidad de que el inspector no encuentre ningún fusible defectuoso. La probabilidad de $x = 0$ es

$$f(0) = \frac{\binom{5}{0}\binom{7}{3}}{\binom{12}{3}} = \frac{\left(\frac{5!}{0!5!}\right)\left(\frac{7!}{3!4!}\right)}{\left(\frac{12!}{3!9!}\right)} = \frac{(1)(35)}{220} = 0.1591$$

Si la probabilidad de cero fusibles defectuosos es $f(0) = 0.1591$, se concluye que la probabilidad de hallar por lo menos un fusible defectuoso debe ser $1 - 0.1591 = 0.8409$. Así, existe una probabilidad razonablemente alta de que el inspector encuentre por lo menos un fusible defectuoso.

La media y la varianza de una distribución hipergeométrica son las siguientes.

$$E(x) = \mu = n\left(\frac{r}{N}\right) \tag{5.13}$$

$$\text{Var}(x) = \sigma^2 = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) \tag{5.14}$$

En el ejemplo anterior $n = 3$, $r = 5$ y $N = 12$. Por tanto, la media y la varianza del número de fusibles defectuosos es

$$\mu = n\left(\frac{r}{N}\right) = 3\left(\frac{5}{12}\right) = 1.25$$

$$\sigma^2 = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) = 3\left(\frac{5}{12}\right)\left(1 - \frac{5}{12}\right)\left(\frac{12-3}{12-1}\right) = 0.60$$

La desviación estándar es $\sigma = \sqrt{0.60} = 0.77$.

NOTAS Y COMENTARIOS

Considere una distribución hipergeométrica con n ensayos. Sea $p = (r/N)$ la probabilidad de un éxito en el primer ensayo. Si el tamaño de la población es grande, el término $(N - n)/(N - 1)$ de la ecuación (5.14) se aproxima a 1. Entonces, el valor esperado y la varianza se expresan como $E(x) = np$ y $\text{Var}(x) = np(1 - p)$. Preste atención a que estas expresiones son las mismas que se usan para calcular el valor esperado y la varianza en una distribución binomial, ecuaciones (5.9) y (5.10). Cuando el tamaño de la población es grande, se aproxima una distribución hipergeométrica mediante una distribución binomial con n ensayos y probabilidad de éxito $p = (r/N)$.

Ejercicios

Métodos



46. Suponga que $N = 10$ y $r = 3$. Calcule las probabilidades hipergeométricas correspondientes a los valores siguiente de n y x.
   a. $n = 4, x = 1$.
   b. $n = 2, x = 2$.
   c. $n = 2, x = 0$.
   d. $n = 4, x = 2$.
47. Suponga que $N = 15$ y $r = 4$. ¿Cuál es la probabilidad de $x = 3$ para $n = 10$?

Aplicaciones

48. En una encuesta realizada por Gallup Organization, se les preguntó a los interrogados, "Cuál es el deporte que prefieres ver". Futbol y básquetbol ocuparon el primero y segundo lugar de preferencia (www.gallup.com, 3 de enero de 2004). Si en un grupo de 10 individuos, siete prefieren futbol y tres prefieren básquetbol. Se toma una muestra aleatoria de tres de estas personas.
   a. ¿Cuál es la probabilidad de que exactamente dos prefieren el futbol?
   b. ¿De que la mayoría (ya sean dos o tres) prefiere el futbol?
49. Blackjack, o veintiuno, como se le suele llamar, es un popular juego de apuestas en los casinos de Las Vegas. A un jugador se le reparten dos cartas. Las figuras (sotas, reinas y reyes) y los 10 valen 10 puntos. Los ases valen 1 u 11. Una baraja de 52 cartas tiene 16 cartas que valen 10 (sotas, reinas, reyes y dieces) y cuatro ases.

a. ¿Cuál es la probabilidad de que las dos cartas repartidas sean ases o cartas que valgan 10 puntos?
b. ¿De que las dos cartas sean ases?
c. ¿De que las dos cartas valgan 10?
d. Un blackjack es una carta de 10 puntos y un as que suman 21. Use sus respuestas a los incisos a, b y c para determinar la probabilidad de que a un jugador se le reparta blackjack. (*Indicación:* El inciso c no es un problema hipergeométrico. Desarrolle su propio razonamiento lógico para combinar las probabilidades hipergeométricas de los incisos a, b y c para responder esta pregunta.)



50. Una empresa fabrica computadoras personales en dos fábricas, una en Texas y la otra en Hawai. La fábrica de Texas tiene 40 empleados; la fábrica de Hawai tiene 20 empleados. A una muestra aleatoria de 20 empleados se le pide que llene un cuestionario sobre prestaciones.
    a. ¿Cuál es la probabilidad de que ninguno de los empleados de la muestra trabaje en la fábrica de Hawai?
    b. ¿De que uno de los empleados de la muestra trabaje en la fábrica de Hawai?
    c. ¿De que dos o más de los empleados de la muestra trabajen en la fábrica de Hawai?
    d. ¿De que nueve de los empleados de la muestra trabajen en la fábrica de Texas?
51. En una revista de encuestas se da información sobre la evaluación a los platillos, la decoración y el servicio de varios de los principales restaurantes de Estados Unidos. En 15 de los mejor evaluados restaurantes de Boston, el costo promedio de una cena, que incluye una bebida y la propina, es \$48.60. Usted va a ir en viaje de negocios a Boston y le gustaría cenar en tres de estos restaurantes. Su empresa le pagará máximo \$50 por cena. Sus conocidos en Boston le han informado que en una tercera parte de estos restaurantes una cena cuesta más de \$50. Suponga que escoge al azar tres de estos restaurantes para ir a cenar.
    a. ¿Cuál es la probabilidad de que el costo de ninguna de las cenas sea mayor a la cantidad que paga su empresa?
    b. ¿De que el costo de una de las cenas sea mayor a la cantidad que paga su empresa?
    c. ¿De que el costo de dos de las cenas sea mayor a la cantidad que paga su empresa?
    d. ¿De que el costo de las tres cenas sea mayor a la cantidad que paga su empresa?
52. En un pedido de 10 artículos hay dos defectuosos y ocho no defectuosos. Para la inspección del pedido se tomará una muestra y se inspeccionará. Si se encuentra un artículo defectuoso todo el pedido de 10 artículos será devuelto.
    a. Si toma una muestra de tres artículos, ¿cuál es la probabilidad de que devuelva el pedido?
    b. Si toma una muestra de cuatro artículos, ¿cuál es la probabilidad de que devuelva el pedido?
    c. Si toma una muestra de cinco artículos, ¿cuál es la probabilidad de que devuelva el pedido?
    d. Si la administración desea que la probabilidad de rechazar un pedido en el que haya dos artículos defectuosos y ocho no defectuosos sea 0.90, ¿de qué tamaño recomienda que sea la muestra?

Resumen

Una variable aleatoria da una descripción numérica de los resultados de un experimento. La distribución de probabilidad de una variable aleatoria describe cómo se reparten las probabilidades entre los valores que toma dicha variable. En toda variable aleatoria discreta, x, su distribución de probabilidad se define mediante una función de probabilidad, que se denota $f(x)$ y la cual da la probabilidad que corresponde a cada valor de la variable aleatoria. Una vez que se ha definido la función de probabilidad, es posible calcular el valor esperado, la varianza y la desviación estándar de la variable aleatoria.

La distribución binomial se usa para determinar la probabilidad de x éxitos en n ensayos, siempre que el experimento satisfaga las propiedades siguientes:

1. El experimento consista en una serie de n ensayos idénticos.
2. En cada ensayo haya dos resultados posibles, uno llamado éxito y el otro fracaso.
3. La probabilidad de un éxito no varíe de un ensayo a otro. Por tanto, la probabilidad de fracaso, $(1 - p)$, tampoco variará de un resultado a otro.
4. Los ensayos sean independientes.

Si se satisfacen estas cuatro propiedades, la probabilidad de x éxitos en n ensayos se determina usando la función de probabilidad binomial. También se presentaron las fórmulas para hallar la media y la varianza de una distribución binomial.

La distribución de Poisson se usa cuando se quiere obtener la probabilidad de x ocurrencias de un evento en un determinado intervalo de tiempo o de espacio. Para que se emplee la distribución de Poisson deben satisfacerse las condiciones siguientes:

1. La probabilidad de una ocurrencia del evento es la misma para cualesquier dos intervalos de la misma longitud.
2. La ocurrencia o no–ocurrencia del evento en un determinado intervalo es independiente de la ocurrencia o no–ocurrencia del evento en cualquier otro intervalo.

En la sección 5.6 se presentó la tercera distribución discreta de probabilidad presentada, la distribución hipergeométrica. Es como la binomial, que se usa para calcular la probabilidad de x éxitos en n ensayos, pero, a diferencia de ésta, la probabilidad de éxito si varía de un ensayo a otro.

Glosario

Variable aleatoria Una descripción numérica del resultado de un experimento.

Variable aleatoria discreta Una variable aleatoria que puede asumir un número finito de valores o un número infinito de valores de una sucesión.

Variable aleatoria continua Ésta toma cualquier valor de un intervalo o de una colección de intervalos.

Distribución de probabilidad Descripción de cómo se distribuyen las probabilidades entre los valores de una variable aleatoria.

Función de probabilidad Se denota $f(x)$ y da la probabilidad de que x tome un determinado valor de una variable aleatoria.

Distribución de probabilidad uniforme discreta Distribución de probabilidad para la cual cada posible valor de la variable aleatoria tienen la misma probabilidad.

Valor esperado Medida de localización central de una variable aleatoria.

Varianza Medida de la variabilidad o dispersión de una variable aleatoria.

Desviación estándar Raíz cuadrada positiva de la varianza.

Experimento binomial Un experimento que tiene cuatro propiedades que se dan al principio de la sección 5.4.

Distribución de probabilidad binomial Distribución de probabilidad da la probabilidad de x éxitos en n ensayos de un experimento binomial.

Función de probabilidad binomial La función usada para calcular las probabilidades binomiales.

Distribución de probabilidad de Poisson Distribución de probabilidad da la probabilidad de x ocurrencias de un evento en un determinado intervalo de tiempo o de espacio.

Función de probabilidad de Poisson La función usada para calcular las probabilidades de Poisson.

Distribución de probabilidad hipergeométrica Distribución de probabilidad da la probabilidad de x éxitos en n ensayos a partir de una población en la que hay r éxitos y $N - r$ fracasos.

Función de probabilidad hipergeométrica La función usada para calcular probabilidades hipergeométricas

Fórmulas clave

Función de probabilidad uniforme discreta

$$f(x) = 1/n \tag{5.3}$$

Valor esperado en una variable aleatoria discreta

$$E(x) = \mu = \Sigma x f(x) \tag{5.4}$$

Varianza en una variable aleatoria discreta

$$\text{Var}(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x) \tag{5.5}$$

Número de resultados experimentales en los que se encuentran exactamente *x* éxitos en *n* ensayos

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{5.6}$$

Función de probabilidad binomial

$$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)} \tag{5.8}$$

Valor esperado en una distribución binomial

$$E(x) = \mu = np \tag{5.9}$$

Varianza en una distribución binomial

$$\text{Var}(x) = \sigma^2 = np(1-p) \tag{5.10}$$

Función de probabilidad de Poisson

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \tag{5.11}$$

Función de probabilidad hipergeométrica

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \qquad \text{para } 0 \leq x \leq r \tag{5.12}$$

Valor esperado en la distribución hipergeométrica

$$E(x) = \mu = n\left(\frac{r}{N}\right) \tag{5.13}$$

Varianza en la distribución hipergeométrica

$$\text{Var}(x) = \sigma^2 = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) \tag{5.14}$$

Ejercicios complementarios

53. El *Barron´s* Big Money Poll preguntó a 131 gerentes de inversiones de Estados Unidos acerca de sus puntos de vista sobre las inversiones a corto plazo (*Barron´s*, 28 de octubre de 2002). De acuerdo con las respuestas 4% se encontraban muy optimistas, 39 % se encontraban optimistas, 29% se encontraban neutrales, 21% se encontraban pesimistas y 7% se encontraban muy pesimistas. Sea x la variable aleatoria que refleje el grado de optimismo y que vaya desde $x = 1$ para muy pesimista hasta $x = 5$ para muy optimista.
    a. Elabore una distribución de probabilidad para el grado de optimismo de los gerentes de inversiones.
    b. Calcule el valor esperado del grado de optimismo.
    c. Calcule la varianza y la desviación estándar del grado de optimismo.
    d. Haga un comentario sobre lo que le dicen sus resultados acerca del grado de optimismo y su variabilidad.

54. La American Association of Individual Investors publica una guía anual con los principales fondos mutualistas (*The Individual Investor´s Guide to the Top Mutual Founds*, 22[a] ed., American Association of Individual Investors, 2003). En la tabla 5.10 se presenta la clasificación de 29 fondos mutualistas de acuerdo con el riesgo.
    a. Sea x una variable que va desde $x = 1$ con el menor riesgo hasta el mayor riesgo con $x = 5$. Elabore una distribución de probabilidad para el nivel de riesgo.
    b. ¿Cuál es el valor esperado y la varianza del nivel de riesgo?
    c. Se encontró que 11 de éstos eran fondos de renta fija. De ellos siete se clasificaron como bajos y cuatro como abajo del promedio. Compare el riesgo de los fondos de renta fija con los 18 fondos de acciones.

TABLA 5.10 DE 29 FONDOS MUTUALISTAS

Número de fondos	Nivel de riesgo: categorías
Bajo	7
Bajo el promedio	6
Promedio	3
Sobre el promedio	6
Alto	7

55. Al hacer el presupuesto de gastos para el próximo año en una universidad, se obtuvieron los siguientes pronósticos de gastos (dados en millones de dólares) $9, $10, $11, $12 y $13. Como no se sabe cuáles son los gastos actuales, a los gastos calculados se les asignaron las probabilidades 0.3, 0.2, 0.25, 0.05 y 0.2.
    a. Dé la distribución de probabilidad para estos pronósticos de gastos.
    b. ¿Cuál es el valor esperado en estos pronósticos de gastos?
    c. ¿Cuál es la varianza en el pronóstico de gastos para el año próximo?
    d. Si las proyecciones de ingreso estiman que éste será de $12 millones, ¿cómo será la situación financiera de la universidad?

56. En un estudio realizado por la Bureau of Transportation Statistics se encontró que, en promedio, la duración del recorrido de la casa al trabajo de una persona es de 26 minutos. También que 5% de las personas necesitan más de una hora para transportarse de su casa al trabajo.
    a. Si interroga a 20 de estas personas, ¿cuál es la probabilidad de que informen que necesitan más de una hora para ir de su casa el trabajo?
    b. Si interroga a 20 de estas personas, ¿cuál es la probabilidad de que ninguna de ellas informe que necesita más de una hora para ir de su casa al trabajo?

c. Si en una empresa hay 2000 empleados, ¿cuál es el número esperado de empleados que necesita más de una hora para trasladarse de su casa al trabajo?
d. Si en una empresa hay 2000 empleados, ¿cuál es la varianza y la desviación estándar del número de empleados que necesitan más de una hora para trasladarse de su casa al trabajo.

57. Una empresa piensa entrevistar a los usuarios de Internet para saber cómo será recibida su página por los grupos de las distintas edades. De acuerdo con la Census Bureau, 40% de las personas entre 18 y 54 años y 12% de las personas de 55 años o más usan Internet.
a. ¿Cuántas personas entre 18 y 54 años hay que contactar para hallar un número esperado de por lo menos 10 usuarios de Internet?
b. ¿Cuántas personas de 55 años o más hay que contactar para hallar un número esperado de por lo menos 10 usuarios de Internet?
c. Si se contacta el número de personas entre 18 y 54 años sugerido por el inciso a, ¿cuál es la desviación estándar del número que será usuario de Internet?
d. Si se contacta el número de personas de entre 55 años o más sugerido por el inciso b, ¿cuál es la desviación estándar del número de quienes serán usuarios de Internet?

58. Muchas empresas usan una técnica de control de calidad conocida como muestreo de aceptación para vigilar los pedidos que reciben de piezas, materia prima, etc. En la industria electrónica, los componentes se suelen recibir por lotes grandes. La inspección de una muestra de *n* componentes se considera como *n* ensayos de un experimento binomial. El resultado de la revisión de cada componente (ensayo) es que el componente sea clasificado como bueno o como defectuoso. Reynolds Electronics acepta el lote de un determinado distribuidor si los componentes defectuosos encontrados en el lote no son más de 1%. Suponga que se prueba una muestra aleatoria de cinco artículos del último lote recibido.
a. Asuma que 1% del lote recibido está defectuoso. Calcule la probabilidad de que ningún elemento de la muestra esté defectuoso.
b. Admita que 1% del lote recibido está defectuoso. Calcule la probabilidad de que exactamente un elemento de la muestra esté defectuoso.
c. ¿Cuál es la probabilidad de encontrar uno o más artículos defectuosos si 1% del lote está defectuoso?
d. ¿Estaría usted tranquilo al aceptar el lote si se encuentra un artículo defectuoso? ¿Por qué sí o por qué no?

59. La tasa de desempleo es 4.1% (*Barron's*, 4 de septiembre de 2000). Suponga que selecciona aleatoriamente 100 personas empleables.
a. ¿Cuál es el número esperado de personas que están desempleadas?
b. ¿Cuál es la varianza y la desviación estándar del número de personas que están desempleadas?

60. Un sondeo de Zogby encontró que de los estadounidenses para quienes la música es "muy importante" en su vida, 30% dice que su estación de radio "siempre" toca la clase de música que le gusta. Suponga que toma una muestra de 800 personas para quienes la música es muy importante en su vida.
a. ¿Cuántas afirmarían que su estación de radio siempre toca la música que les gusta?
b. ¿Cuál es la desviación estándar del número de interrogados para quienes su estación de radio siempre toca la música que les gusta?
c. ¿Cuál es la desviación estándar del número de interrogados para quienes su estación de radio no siempre toca la música que les gusta?

61. A un lavado de coches los automóviles llegan en forma aleatoria e independiente; la probabilidad de una llegada es la misma en cualesquiera dos intervalos de la misma duración. La tasa de llegada media es 15 automóviles por hora. ¿Cuál es la probabilidad de que en una hora cualquiera de operación lleguen 20 o más automóviles?

62. En un proceso nuevo de producción automática hay en promedio 1.5 interrupciones por día. Debido al elevado costo de las interrupciones, los directivos están preocupados por la posibilidad de que en un día haya tres o más interrupciones. Suponga que las interrupciones se presentan en forma aleatoria, que la probabilidad de una interrupción es la misma en cualesquiera dos intervalos de una misma duración y que las interrupciones en un intervalo de tiempo son independientes de

las interrupciones en otro intervalo de tiempo. ¿Cuál es la probabilidad de que haya tres o más interrupciones en un día?

63. Un director regional responsable del desarrollo de los negocios en una determinada área está preocupado por el número de fracasos de pequeños negocios. Si en promedio fracasan 10 pequeños negocios por mes, ¿Cuál es la probabilidad de que exactamente cuatro pequeños negocios fracasen en un mes determinado? Suponga que la probabilidad de fracasos es la misma en cada dos meses que se tomen y que la ocurrencia o no–ocurrencia de fracasos en un determinado mes es independiente de la ocurrencia o no–ocurrencia de fracasos en cualquier otro mes

64. Las llegadas de los clientes a un banco son aleatorias e independientes; la probabilidad de una llegada en un lapso cualquiera de un minuto es la misma que la probabilidad de una llegada en otro lapso cualquiera de un minuto. Conteste las preguntas siguientes suponiendo que la tasa media de llegadas en un lapso de un minuto es tres clientes.
   a. ¿Cuál es la probabilidad de exactamente tres llegadas en un minuto?
   b. ¿Cuál es la probabilidad de por lo menos tres llegadas en un minuto?

65. Una baraja contiene 52 cartas, de las cuales cuatro son ases. ¿Cuál es la probabilidad de que en una repartición de cinco cartas haya:
   a. Un par de ases?
   b. Exactamente un as?
   c. Ningún as?
   d. Por lo menos un as?

66. En la semana que terminó el 16 de septiembre de 2001, Tiger Woods estuvo a la cabeza en ganancia de dinero en el PGA Tour, con una ganancia total de $5 517 777. De los 10 principales jugadores en ganancias de dinero siete usaron pelotas de golf de la marca Titleist (www.pgatour.com). Suponga que toma al azar a dos de estos principales ganadores.
   a. ¿Cuál es la probabilidad de que exactamente uno use una pelota de golf de la marca Titleist?
   b. ¿De que los dos usen una pelota de golf de la marca Titleist?
   c. ¿De que ninguno use una pelota de golf de la marca Titleist?

Apéndice 5.1 Distribuciones de probabilidad con Minitab

Los paquetes para estadística como Minitab ofrecen procedimientos relativamente fáciles y eficientes para calcular probabilidades binomiales. En este apéndice se muestra paso a paso el procedimiento para hallar las probabilidades binomiales del problema de la tienda de ropa Martin Clothing Store de la sección 5.4. Recuerde que las probabilidades binomiales deseadas son para $n = 10$ y $p = 0.30$. Antes de empezar con la rutina de Minitab, el usuario debe ingresar los valores deseados de la variable aleatoria en una columna de la hoja de cálculo. Aquí se han ingresado los valores 0, 1, 2, . . ., 10 en la columna 1 (véase la figura 5.5) para generar la distribución de probabilidad binomial completa. Los pasos para obtener las probabilidades binomiales deseadas usando Minitab son los siguientes.

Paso 1. Seleccionar el menú **Calc**
Paso 2. Elegir **Probability distributions**
Paso 3. Elegir **Binomial**
Paso 4. Cuando aparezca el cuadro de diálogo Binomial Distribution:
Seleccionar **Probability**
Ingresar 10 en el cuadro **Number of trials**
Ingresar 0.3 en el cuadro **Probability of succes**
Ingresar C1 en el cuadro **Input column**
Clic en **OK**

El resultado que da Minitab con las probabilidades binomiales aparecerá como se muestra en la figura 5.5.

De manera similar, Minitab proporciona probabilidades de Poisson e hipergeométricas. Por ejemplo, para calcular probabilidades de Poisson, las únicas diferencias están en el paso 3, en el que se deberá seleccionar la opción **Poisson** y en el paso 4, en el que se deberá ingresar **Mean** en lugar del número de ensayos y la probabilidad de éxito

Apéndice 5.2 Distribuciones de probabilidad discreta con Excel

Excel proporciona funciones para calcular las probabilidades de las distribuciones binomial, de Poisson e hipergeométrica tratadas en este capítulo. La función de Excel para calcular probabilidades binomiales es DISTR.BINOM. Esta función tiene cuatro argumentos: x (el número de éxitos), n (el número de ensayos), p (la probabilidad de éxito) y acumulado. Se usa FALSO como cuarto argumento (acumulado) si se quiere la probabilidad de x éxitos y VERDADERO se usa como cuarto argumento si se desea la probabilidad acumulada de x o menos éxitos. A continuación se muestra cómo calcular la probabilidad de 0 a 10 éxitos en el caso del problema de la tienda de ropa Martin Clothing Store de la sección 5.4 (véase figura 5.5).

A medida que se describe la elaboración de la hoja de cálculo consulte la figura 5.6; la hoja de cálculo con las fórmulas aparece en segundo plano y la hoja de cálculo con los valores en primer plano. En la celda B1 ingrese el número de ensayos (10), en la celda B2 la probabilidad de

FIGURA 5.6 HOJA DE CÁLCULO DE EXCEL PARA CALCULAR PROBABILIDADES BINOMIALES

	A	B	C	D
1	**Number of Trials (*n*)**	10		
2	**Probability of Success (*p*)**	0.3		
3				
4		***x***	***f*(*x*)**	
5		0	=BINOMDIST(B5,$B$1,$B$2,FALSE)	
6		1	=BINOMDIST(B6,$B$1,$B$2,FALSE)	
7		2	=BINOMDIST(B7,$B$1,$B$2,FALSE)	
8		3	=BINOMDIST(B8,$B$1,$B$2,FALSE)	
9		4	=BINOMDIST(B9,$B$1,$B$2,FALSE)	
10		5	=BINOMDIST(B10,$B$1,$B$2,FALSE)	
11		6	=BINOMDIST(B11,$B$1,$B$2,FALSE)	
12		7	=BINOMDIST(B12,$B$1,$B$2,FALSE)	
13		8	=BINOMDIST(B13,$B$1,$B$2,FALSE)	
14		9	=BINOMDIST(B14,$B$1,$B$2,FALSE)	
15		10	=BINOMDIST(B15,$B$1,$B$2,FALSE)	
16				

	A	B	C	D
1	**Number of Trials (*n*)**	10		
2	**Probability of Success (*p*)**	0.3		
3				
4		***x***	***f*(*x*)**	
5		0	0.0282	
6		1	0.1211	
7		2	0.2335	
8		3	0.2668	
9		4	0.2001	
10		5	0.1029	
11		6	0.0368	
12		7	0.0090	
13		8	0.0014	
14		9	0.0001	
15		10	0.0000	
16				

éxito y en las celdas B5:B15 los valores de la variable aleatoria. Con los pasos siguientes generará las probabilidades deseadas.

Paso 1. Usar la función DISTR.BINOM para calcular la probabilidad de $x = 0$ ingresando la fórmula siguiente en la celda C5:

=BINOMDIST(B5,$B$1,$B$2,FALSO)

Paso 2. Copiar la fórmula de la celda C5 en las celdas C6:C15.

La hoja de cálculo con los valores en la figura 5.6 muestra que las probabilidades obtenidas son las mismas que aparecen en la figura 5.5. Las probabilidades de Poisson e hipergeométrica se calculan de manera similar. Se emplean las funciones POISSON y DISTR.HIPERGEOM. La herramienta de Excel Insertar función puede ayudar al usuario a ingresar los argumentos adecuados para estas funciones (véase apéndice 2.2).

CAPÍTULO 6



Distribuciones de probabilidad continua

CONTENIDO



LA ESTADÍSTICA *en* LA PRÁCTICA

PROCTER & GAMBLE*
CINCINNATI, OHIO

Procter & Gamble (P&G) produce y comercializa productos como detergentes, pañales desechables, productos farmacéuticos que no requieren receta, dentífricos, jabones de tocador y toallas de papel. En todo el mundo P&G tiene la marca líder en más categorías que cualquiera otra empresa de productos de consumo. Desde su fusión con Gillette, P&G también comercializa rasuradoras, navajas para afeitar y muchos otros productos para el cuidado personal.

Al ser uno de los líderes en aplicación de los métodos estadísticos para la toma de decisiones, P&G emplea personas con diversas formaciones académicas: ingenieros, especialistas en estadística, en investigación de operaciones y en negocios. Las principales tecnologías cuantitativas en las que estos profesionistas aplican sus conocimientos son decisiones probabilísticas y análisis de riesgos, simulación avanzada, mejoramiento de la calidad y métodos cuantitativos (por ejemplo, programación lineal, análisis de regresión, análisis de probabilidad).

La División de Productos Químicos para la Industria de P&G es una de las principales proveedoras de alcoholes grasos obtenidos de sustancias naturales, como el aceite de coco, y de derivados del petróleo. La división deseaba saber qué riesgos económicos y cuáles oportunidades existen para la expansión de sus instalaciones dedicadas a la producción de alcoholes grasos; por tanto, solicitó la ayuda de los expertos de P&G en decisiones probabilísticas y en análisis de riesgos. Después de estructurar y modelar el problema, los expertos determinaron que la clave para la rentabilidad era la diferencia entre los costos de las materias primas provenientes del petróleo y del coco. Los costos futuros no se podían saber, pero los analistas los calcularon mediante las siguientes variables aleatorias continuas.

x = precio del aceite de coco por libra de alcoholes grasos

y

y = precio de la materia prima proveniente del petróleo por libra de alcoholes grasos



Algunos de los muchos productos de Procter & Gamble son bien conocidos. © AFP/Getty Images.

Como la clave de la rentabilidad era la diferencia entre estas dos variables aleatorias, se empleó una tercera variable aleatoria para el análisis $d = x - y$. Para determinar las distribuciones de probabilidad de x y y entrevistaron a varios expertos. Después, esta información se empleó para elaborar una distribución de probabilidad de la diferencia entre los precios d. En esta distribución de probabilidad continua se encontró que la probabilidad de que la diferencia entre los precios fuera \$0.0655 o menos, era 0.90 y que la probabilidad de que la diferencia entre los precios fuera \$0.035 o menos era 0.50. Además, la probabilidad de que la diferencia fuera \$0.0045 o menos era sólo 0.10.[†]

La dirección de esta división pensó que la clave para alcanzar un consenso estaba en poder cuantificar el impacto de las diferencias entre los precios de las materias primas. Las probabilidades obtenidas se usaron en un análisis sensible a la diferencia entre los precios de las materias primas. Este análisis arrojó suficiente información como para sustentar una recomendación para los directivos.

Usar variables aleatorias continuas y sus distribuciones de probabilidad ayudó a P&G a analizar los riesgos económicos relacionados con su producción de alcoholes grasos. En este capítulo el lector conocerá las variables aleatorias continuas y sus distribuciones de probabilidad, entre ellas una de las distribuciones de probabilidad más importantes en la estadística, la distribución normal.

---

*Los autores agradecen a Joel Kahn de P&G por proporcionar este artículo para *La estadística en la práctica.*

†Las diferencias de precios dadas aquí están modificadas para proteger los datos.

En el capítulo anterior se estudiaron las variables aleatorias discretas y sus distribuciones de probabilidad. En este capítulo se tratan las variables aleatorias continuas. En específico verá tres distribuciones de probabilidad continua: la uniforme, la normal y la exponencial.

Una diferencia fundamental entre las variables aleatorias discretas y las variables aleatorias continuas es cómo se calculan las probabilidades. En las variables aleatorias discretas la función de probabilidad $f(x)$ da la probabilidad de que la variable aleatoria tome un valor determinado. En las variables aleatorias continuas, la contraparte de la función de probabilidad es la **función de densidad de probabilidad**, que también se denota $f(x)$. La diferencia está en que la función de densidad de probabilidad no da probabilidades directamente. Si no que el área bajo la curva de $f(x)$ que corresponde a un intervalo determinado proporciona la probabilidad de que la variable aleatoria tome uno de los valores de ese intervalo. De manera que cuando se calculan probabilidades de variables aleatorias continuas se calcula la probabilidad de que la variable aleatoria tome alguno de los valores dentro de un intervalo.

Como en cualquier punto determinado el área bajo la gráfica de $f(x)$ es cero, una de las consecuencias de la definición de la probabilidad de una variable aleatoria continua es que la probabilidad de cualquier valor determinado de la variable aleatoria es cero. Estos conceptos se demuestran en la sección 6.1 con una variable que tiene una distribución uniforme.

Gran parte del capítulo se dedica a describir y mostrar aplicaciones de la distribución normal. La distribución normal es muy importante por tener muchas aplicaciones y un amplio uso en la inferencia estadística. El capítulo concluye con el estudio de la distribución exponencial. La distribución exponencial es útil en aplicaciones en las que intervienen factores como tiempos de espera y tiempos de servicios.



Distribución de probabilidad uniforme

Siempre que una probabilidad sea proporcional a la longitud del intervalo, la variable aleatoria estará distribuida uniformemente.

Considere una variable aleatoria x que representa el tiempo de vuelo de un avión que viaja de Chicago a Nueva York. Suponga que el tiempo de vuelo es cualquier valor en el intervalo de 120 minutos a 140 minutos. Dado que la variable aleatoria x toma cualquier valor en este intervalo, x es una variable aleatoria continua y no una variable aleatoria discreta. Admita que cuenta con datos suficientes como para concluir que la probabilidad de que el tiempo de vuelo esté en cualquier intervalo de 1 minuto es el mismo que la probabilidad de que el tiempo de vuelo esté en cualquier otro intervalo de 1 minuto dentro del intervalo que va de 120 a 140 minutos. Como cualquier intervalo de 1 minuto es igual de probable, se dice que la variable aleatoria x tiene una **distribución de probabilidad uniforme**. La función de densidad de probabilidad que define la distribución uniforme de la variable aleatoria tiempo de vuelo, es

$$f(x) = \begin{cases} 1/20 & \text{para } 120 \le x \le 140 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

La figura 6.1 es una gráfica de esta función de densidad de probabilidad. En general, la función de densidad de probabilidad uniforme de una variable aleatoria x se define mediante la fórmula siguiente.

FUNCIÓN DE DENSIDAD DE PROBABILIDAD UNIFORME

$$f(x) = \begin{cases} \dfrac{1}{b - a} & \text{para } a \le x \le b \\ 0 & \text{en cualquier otro caso} \end{cases} \tag{6.1}$$

En el caso de la variable aleatoria tiempo de vuelo, $a = 120$ y $b = 140$.

FIGURA 6.1 DISTRIBUCIÓN DE PROBABILIDAD UNIFORME PARA EL TIEMPO DE VUELO

$f(x)$

$\frac{1}{20}$

120 125 130 135 140 x

Tiempo de vuelo en minutos

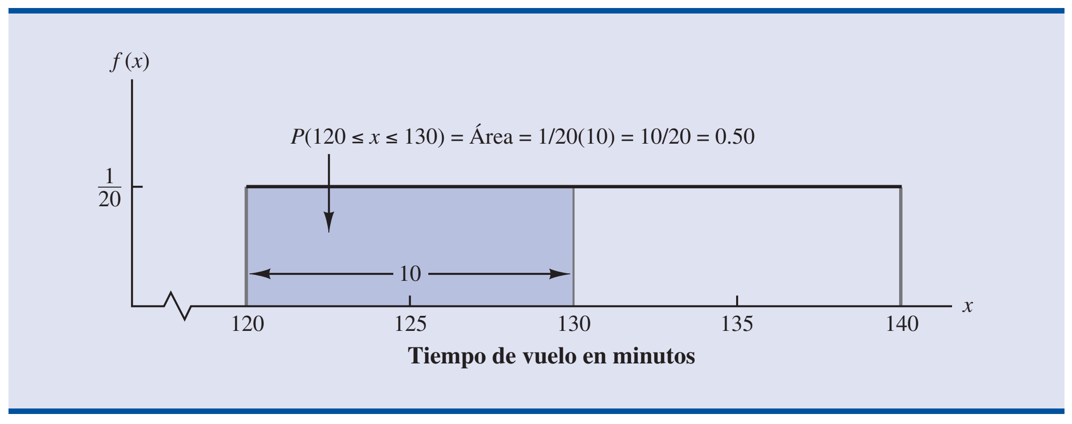
Como se hizo notar en la introducción, en el caso de una variable aleatoria continua, sólo se considera la probabilidad en términos de la posibilidad de que la variable aleatoria tome un valor dentro de un determinado intervalo. En el ejemplo del tiempo de vuelo, una pregunta aceptable acerca de una probabilidad es: ¿Cuál es la probabilidad de que el tiempo de vuelo se encuentre entre 120 y 130 minutos? Es decir, ¿cuál es $P(120 \leq x \leq 130)$? Como el tiempo de vuelo debe estar entre 120 y 140 minutos y como se ha dicho que la probabilidad es uniforme en este intervalo, es factible decir que $P(120 \leq x \leq 130) = 0.50$. En la sección siguiente se muestra que esta probabilidad se calcula como el área bajo la gráfica de $f(x)$ desde 120 hasta 130 (véase figura 6.2)

Áreas como medida de probabilidad

Ahora una observación acerca de la gráfica de la figura 6.2. Considere el área bajo la gráfica de $f(x)$ en el intervalo que va de 120 a 130. Esta área es rectangular y el área de un rectángulo es simplemente el ancho multiplicado por la altura. Si el ancho del intervalo es igual a $130 - 120 = 10$ y la altura es igual al valor de la función de densidad de probabilidad $f(x) = 1/20$, se tiene, área = ancho $\times$ alto = $10(1/20) = 10(1/20) = 10/20 = 0.50$.



FIGURA 6.2 EL ÁREA PROPORCIONA LA PROBABILIDAD DE QUE EL TIEMPO DE VUELO ESTÉ ENTRE 120 Y 130 MINUTOS.



¿Qué observación se puede hacer acerca del área bajo la curva de $f(x)$ y la probabilidad? ¡Son idénticas! En efecto, esta observación es correcta y válida para todas las variables aleatorias continuas. Una vez que se ha dado la función de densidad de probabilidad $f(x)$, la probabilidad de que x tome un valor entre algún valor menor x_1 y otro valor mayor x_2 se encuentra calculando el área bajo la gráfica de $f(x)$ y sobre el intervalo de va de x_1 a x_2.

Dada la distribución uniforme del tiempo de vuelo y usando la interpretación de área como probabilidad es posible contestar cualquier pregunta acerca de la probabilidad de los tiempos de vuelo. Por ejemplo, ¿cuál es la probabilidad de un tiempo de vuelo entre 128 y 136 minutos? El ancho del intervalo es $136 - 128 = 8$. Como la altura uniforme de $f(x) = 1/120$, se ve que $P(128 \leq x \leq 136) = 8(1/20) = 0.40$.

Observe que $P(120 \leq x \leq 140) = 20(1/20) = 1$; es decir, el área total bajo la gráfica de $f(x)$ es igual a 1. Esta propiedad es válida para todas las distribuciones de probabilidad continua y es el análogo de la condición de que la suma de las probabilidades debe ser igual a 1 en el caso de una función de probabilidad discreta.

Dos diferencias importantes sobresalen entre el tratamiento de una variable aleatoria continua y el tratamiento de una variable aleatoria discreta.

Para ver que la probabilidad de un solo punto es 0, consulte la figura 6.2 y calcule la probabilidad de un solo punto, por ejemplo, $x = 125$. $P(x = 125) = P(125 \leq x \leq 125) = 0(1/20) = 0$.

1. Ya no se habla de la probabilidad de que una variable aleatoria tome un determinado valor. Se habla de la probabilidad de que una variable aleatoria tome un valor dentro de un intervalo dado.
2. La probabilidad de que una variable aleatoria continua tome un valor dentro de un determinado intervalo que va de x_1 a x_2 se define como el área bajo la gráfica de la función de densidad de probabilidad entre x_1 y x_2. Como un solo punto es un intervalo cuyo ancho es cero, esto implica que la probabilidad de que una variable aleatoria continua tome un valor exacto, cualquiera, es cero. Esto también significa que en cualquier intervalo la probabilidad de que una variable aleatoria continua tome un valor es la misma, ya sea que se incluyan o no los extremos del intervalo.

El cálculo del valor esperado y de la varianza de una variable aleatoria continua es análogo al de una variable aleatoria discreta. Sin embargo, como en este caso interviene el cálculo integral la deducción de estas fórmulas queda para cursos más avanzados.

En el caso de la distribución de probabilidad continua uniforme presentada en esta sección, las fórmulas para el valor esperado y para la varianza son

$$E(x) = \frac{a + b}{2}$$

$$\text{Var}(x) = \frac{(b - a)^2}{12}$$

En estas fórmulas a es el menor valor y b es el mayor valor que toma la variable aleatoria.

Al aplicar estas fórmulas a la distribución uniforme de los tiempos de vuelo de Chicago a Nueva York, se obtiene,

$$E(x) = \frac{(120 + 140)}{2} = 130$$

$$\text{Var}(x) = \frac{(140 - 120)^2}{12} = 33.33$$

La desviación estándar de los tiempos de vuelo se encuentra sacando la raíz cuadrada de la varianza. Por tanto, $\sigma = 5.77$ minutos.

NOTAS Y COMENTARIOS

Para ver más claramente por qué la altura de una función de densidad de probabilidad no es una probabilidad, considere la variable aleatoria cuya distribución de probabilidad uniforme es la siguiente.

$$f(x) = \begin{cases} 2 & \text{para } 0 \leq x \leq 0.5 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

La altura de la función de densidad de probabilidad, $f(x)$ es 2 para todos los valores de x entre 0 y 0.5. Pero se sabe que las probabilidades nunca pueden ser mayores a 1. Por tanto, $f(x)$ no se interpreta como la probabilidad de x.

Ejercicios

Métodos



1. La variable aleatoria x está distribuida uniformemente entre 1.0 y 1.5.
   a. Dé la gráfica de la función de densidad de probabilidad.
   b. Calcule $P(x = 1.25)$.
   c. Calcule $P(1.0 \leq x \leq 1.25)$.
   d. Calcule $P(1.20 < x < 1.5)$.
2. La variable aleatoria x está distribuida uniformemente entre 10 y 20.
   a. Dé la gráfica de la función de densidad de probabilidad.
   b. Calcule $P(x < 15)$.
   c. Calcule $P(12 \leq x \leq 18)$.
   d. Calcule $E(x)$.
   e. Calcule $\text{Var}(x)$.

Aplicaciones

3. En su vuelo de Cincinati a Tampa, Delta Airlines da como tiempo de vuelo 2 horas, 5 minutos. En realidad los tiempos de vuelo están distribuidos uniformemente entre 2 horas y 2 horas, 20 minutos.
   a. Dé la gráfica de la función de densidad de probabilidad del tiempo de vuelo.
   b. ¿Cuál es la probabilidad de que un vuelo no se retrase más de 5 minutos?
   c. ¿De que un vuelo no se retrase más de 10 minutos?
   d. ¿Cuál es el tiempo de vuelo esperado?



4. La mayoría de los lenguajes de computadora tienen una función para generar números aleatorios. En Excel, la función ALEATORIO se usa para generar números aleatorios entre 0 y 1. Si x denota un número aleatorio generado mediante ALEATORIO, entonces x es una variable aleatoria continua, cuya función de densidad de probabilidad es la siguiente.

$$f(x) = \begin{cases} 1 & \text{para } 0 \leq x \leq 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

   a. Haga la gráfica de la función de densidad de probabilidad.
   b. ¿Cuál es la probabilidad de generar un número aleatorio entre 0.25 y 0.75?
   c. ¿De generar un número aleatorio menor o igual que 0.30?
   d. ¿De generar un número aleatorio mayor o igual que 0.60?
   e. Genere 50 números aleatorios ingresando = ALEATORIO() en 50 celdas de una hoja de cálculo de Excel.
   f. Calcule la media y la desviación estándar de los números del inciso e.

5. La *driving distance* de los 100 mejores golfistas del Tour PGA está entre 284.7 y 310.6 yardas (*Golfweek*, 29 de marzo de 2003). Suponga que las *driving distance* de estos golfistas se encuentran uniformemente distribuidas en este intervalo.
   a. Dé una expresión matemática de la función de densidad de probabilidad correspondiente a estas *driving distance*
   b. ¿Cuál es la probabilidad de que la *driving distance* de uno de estos golfistas sea menor que 290 yardas?
   c. ¿De que la *driving distance* de uno de estos golfistas sea por lo menos de 300 yardas?
   d. ¿De que la *driving distance* de uno de estos golfistas esté entre 290 y 305 yardas?
   e. ¿Cuántos de estos jugadores lanzan la pelota por lo menos a 290 yardas?
6. En las botellas de un detergente líquido se indica que el contenido es de 12 onzas por botella. En la operación de producción se llenan las botellas uniformemente de acuerdo con la siguiente función de densidad de probabilidad.

$$f(x) = \begin{cases} 8 & \text{para } 11.975 \leq x \leq 12.100 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

   a. ¿Cuál es la probabilidad de que el contenido de una botella esté entre 12 y 12.05 onzas?
   b. ¿De que el contenido de una botella sea 12.02 onzas o más?
   c. En el control de calidad se acepta que una botella sea llenada con más o menos 0.02 onzas de lo indicado en la etiqueta. ¿Cuál es la probabilidad de que una de las botellas de detergente no satisfaga estos estándares?
7. Suponga que quiere comprar un terreno y sabe que también hay otros compradores interesados.* El vendedor revela que aceptará la oferta mayor que sea superior a $10 000. Si la oferta del competidor x es una variable aleatoria que está uniformemente distribuida entre $10 000 y $15 000.
   a. Asuma que usted ofrece $12 000. ¿Cuál es la probabilidad de que su oferta sea aceptada?
   b. Si usted ofrece $14 000. ¿Cuál es la probabilidad de que su oferta sea aceptada?
   c. ¿Cuál es la cantidad que deberá ofrecer para maximizar la probabilidad de obtener la propiedad?
   d. Suponga que conoce a quien está dispuesto a pagar $16 000 por la propiedad. ¿Consideraría la posibilidad de ofrecer una cantidad menor que la del inciso c?



6.2 Distribución de probabilidad normal

Abraham de Moivre, un matemático francés, publicó en 1733 Doctrina de las posibilidades*. De Moivre dedujo la distribución normal.*

La distribución de probabilidad más usada para describir variables aleatorias continuas es la **distribución de probabilidad normal**. La distribución normal tiene gran cantidad de aplicaciones prácticas, en las cuales la variable aleatoria puede ser el peso o la estatura de las personas, puntuaciones de exámenes, resultados de mediciones científicas, precipitación pluvial u otras cantidades similares. La distribución normal también tiene una importante aplicación en inferencia estadística, tema principal del resto de este libro. En estas aplicaciones, la distribución normal describe qué tan probables son los resultados obtenidos de un muestreo

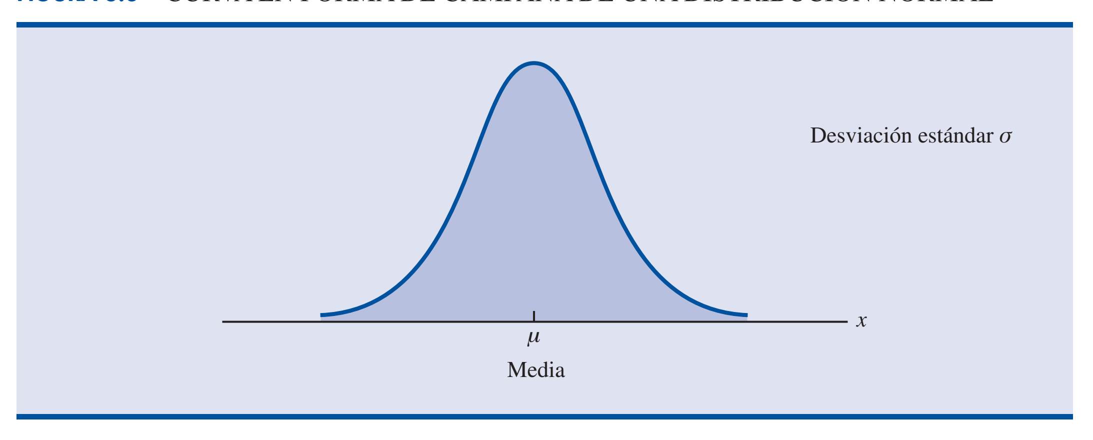
Curva normal

En la figura 6.3 aparece la forma de la distribución normal, una curva normal en forma de campana. A continuación se presenta la función de densidad de probabilidad que define la curva en forma de campana de la distribución normal.

---

* Este ejercicio está basado en un problema sugerido por el profesor Roger Myerson de la Universidad de Northwestern.

FIGURA 6.3 CURVA EN FORMA DE CAMPANA DE UNA DISTRIBUCIÓN NORMAL



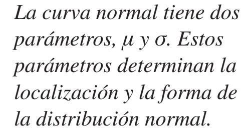
FUNCIÓN DE DENSIDAD DE PROBABILIDAD NORMAL

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad \textbf{(6.2)}$$

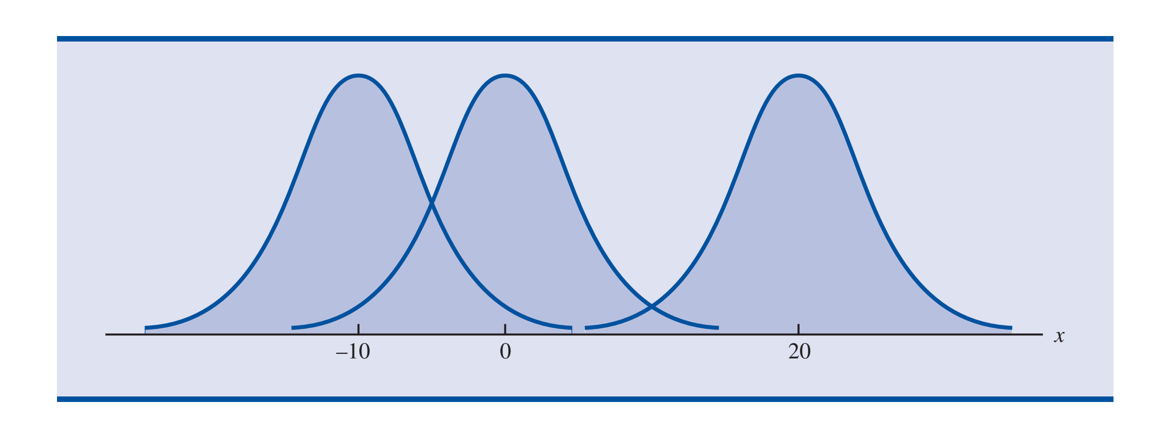
donde

$$\begin{aligned} \mu &= \text{media} \\ \sigma &= \text{desviación estándar} \\ \pi &= 3.14159 \\ e &= 2.71828 \end{aligned}$$

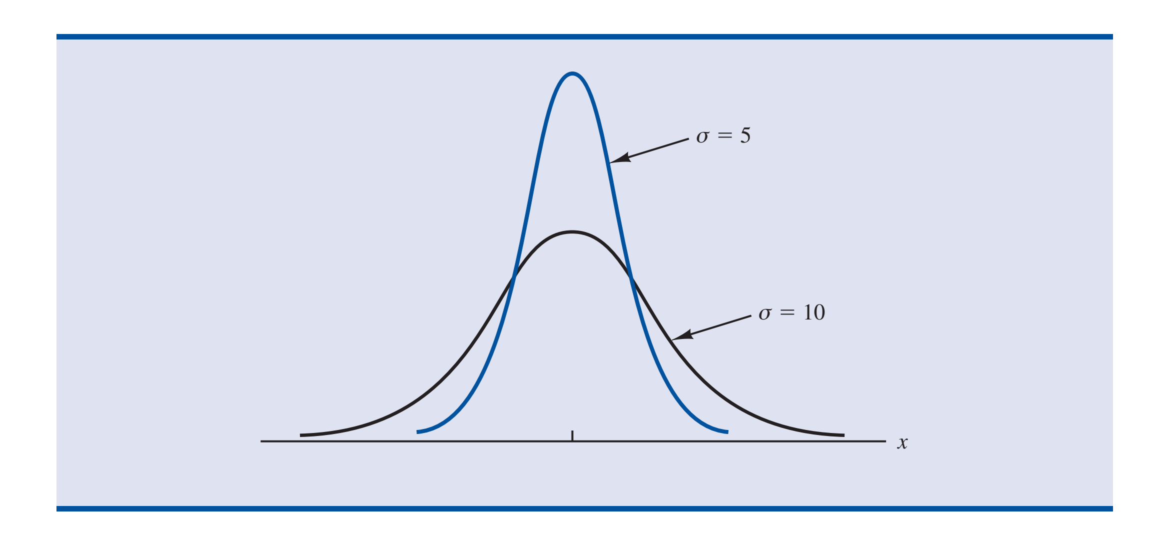
Las siguientes son observaciones importantes acerca de las características de las distribuciones normales.


La curva normal tiene dos parámetros, μ y σ. Estos parámetros determinan la localización y la forma de la distribución normal.

1. Toda la familia de distribuciones normales se diferencia por medio de dos parámetros: la media μ y la desviación estándar σ.
2. El punto más alto de una curva normal se encuentra sobre la media, la cual coincide con la mediana y la moda.
3. La media de una distribución normal puede tener cualquier valor: negativo, positivo o cero. A continuación se muestran tres distribuciones normales que tienen la misma desviación estándar, pero diferentes medias. (−10, 0 y 20).



4. La distribución normal es simétrica, siendo la forma de la curva normal al lado izquierdo de la media, la imagen especular de la forma al lado derecho de la media. Las colas de la curva normal se extienden al infinito en ambas direcciones y en teoría jamás tocan el eje horizontal. Dado que es simétrica, la distribución normal no es sesgada; su sesgo es cero.
5. La desviación estándar determina qué tan plana y ancha es la curva normal. Desviaciones estándar grandes corresponden a curvas más planas y más anchas, lo cual indica mayor variabilidad en los datos. A continuación se muestran dos curvas normales que tienen la misma media pero distintas desviaciones estándar.



6. Las probabilidades correspondientes a la variable aleatoria normal se dan mediante áreas bajo la curva normal. Toda el área bajo la curva de una distribución normal es 1. Como esta distribución es simétrica, el área bajo la curva y a la izquierda de la media es 0.50 y el área bajo la curva y a la derecha de la media es 0.50.
7. Los porcentajes de los valores que se encuentran en algunos intervalos comúnmente usados son:
   a. 68.3% de los valores de una variable aleatoria normal se encuentran más o menos una desviación estándar de la media.
   b. 95.4% de los valores de una variable aleatoria normal se encuentran más o menos dos desviaciones estándar de la media.
   c. 99.7% de los valores de una variable aleatoria normal se encuentran más o menos tres desviaciones estándar de la media.

Estos porcentajes son la base de la regla empírica que se presentó en la sección 3.3.

En la figura 6.4 aparece una gráfica de las propiedades a, b y c.

Distribución de probabilidad normal estándar

Una variable aleatoria que tiene una distribución normal con una media cero y desviación estándar de uno tiene una **distribución normal estándar**. Para designar esta variable aleatoria normal se suele usar la letra z. La figura 6.5 es la gráfica de la distribución normal estándar. Esta distribución tiene el mismo aspecto general que cualquier otra distribución normal, pero tiene las propiedades especiales, $\mu = 0$ y $\sigma = 1$.

FIGURA 6.4 ÁREAS BAJO LA CURVA DE CUALQUIER DISTRIBUCIÓN NORMAL

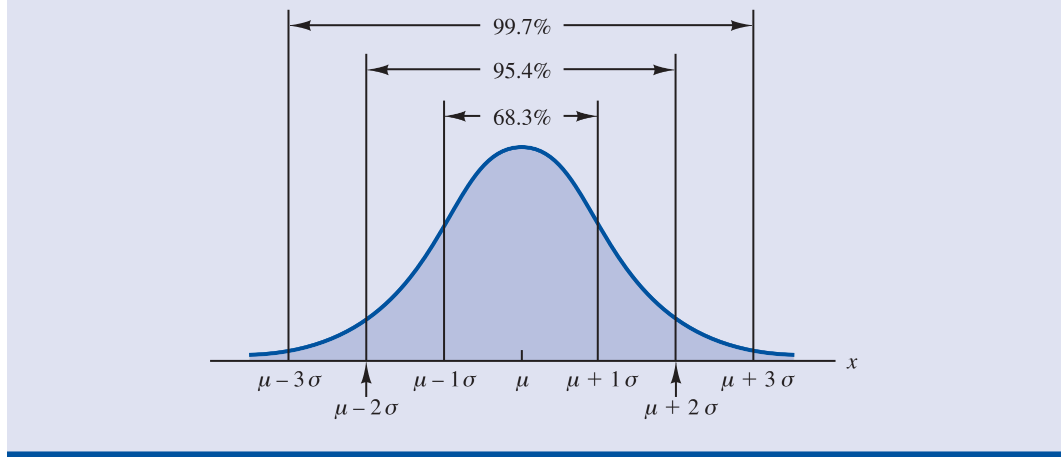
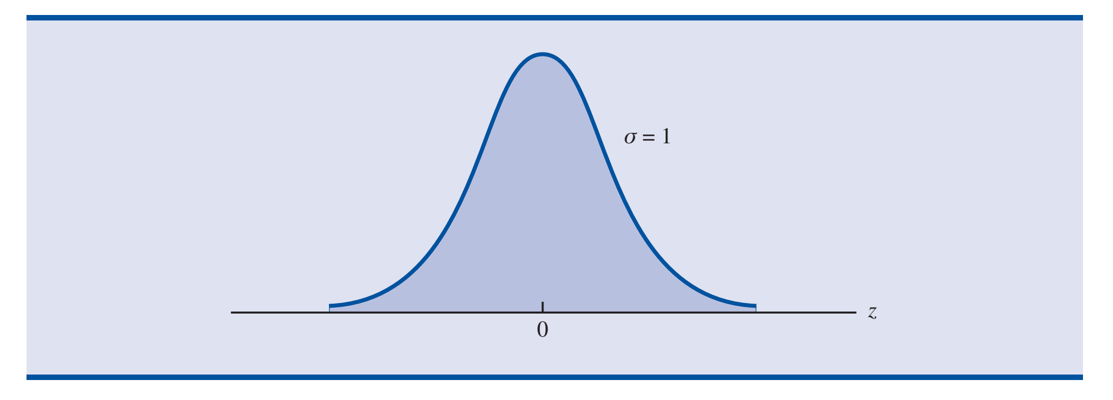


FIGURA 6.5 DISTRIBUCIÓN NORMAL ESTÁNDAR



Dado que $\mu = 0$ y $\sigma = 1$, la fórmula de la función de densidad de probabilidad normal estándar es una versión más simple de la ecuación (6.2).

FUNCIÓN DE DENSIDAD NORMAL ESTÁNDAR

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Como ocurre con otras variables aleatorias continuas, los cálculos de la probabilidad en cualquier distribución normal se hacen calculando el área bajo la gráfica de la función de densidad de probabilidad. Por tanto, para hallar la probabilidad de que una variable aleatoria normal esté dentro de un determinado intervalo, se tiene que calcular el área que se encuentra bajo la curva normal y sobre ese intervalo.

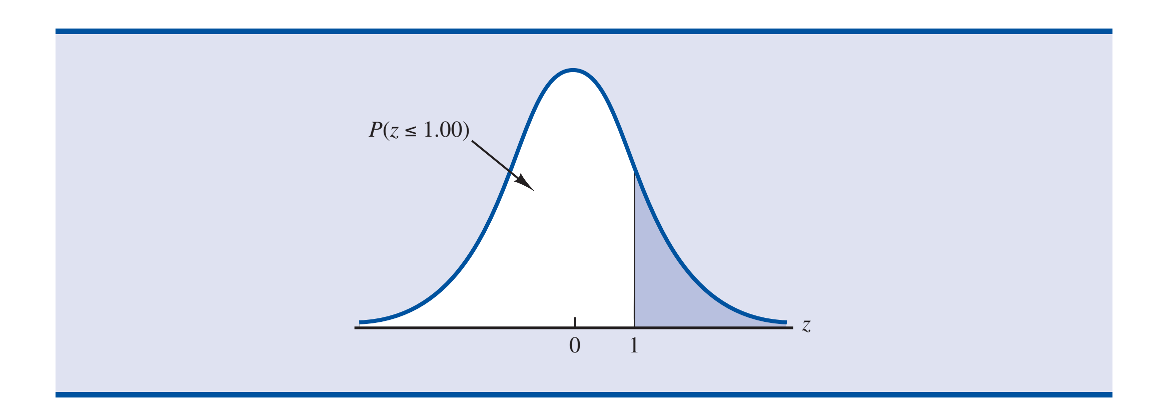
En la función de densidad de probabilidad normal, la altura de la curva varía y para calcular las áreas que representan las probabilidades se requiere de matemáticas más avanzadas.

Para la distribución normal estándar ya se encuentran calculadas las áreas bajo la curva normal y se cuenta con tablas que dan estas áreas y que se usan para calcular las probabilidades. Estas tablas se encuentran en los forros interiores al inicio del libro. La tabla del forro izquierdo contiene áreas, o probabilidades acumuladas, correspondientes a valores de z menores o iguales a la media, cero. La tabla siguiente contiene áreas, o probabilidades acumuladas, correspondientes a valores de z mayores o iguales a la media de cero.

Los tres tipos de probabilidades que se necesitan calcular son: (1) la probabilidad de que la variable aleatoria normal estándar z sea menor o igual que un valor dado; (2) la probabilidad de que z esté entre dos valores dados, y (3) la probabilidad de que z sea mayor o igual que un valor dado. Para mostrar el uso de las tablas de probabilidad acumulada de la distribución normal estándar en el cálculo de estos tres tipos de probabilidades, se consideran algunos ejemplos.

Debido a que la variable aleatoria normal estándar es continua, $P(z \leq 1.00) = P(z < 1.00)$.

Se empieza por mostrar cómo se calcula la probabilidad de que z sea menor o igual a 1.00; es decir $P(z \leq 1.00)$. Esta probabilidad acumulada es el área bajo la curva normal a la izquierda de $z = 1.00$ como se muestra en la gráfica siguiente.

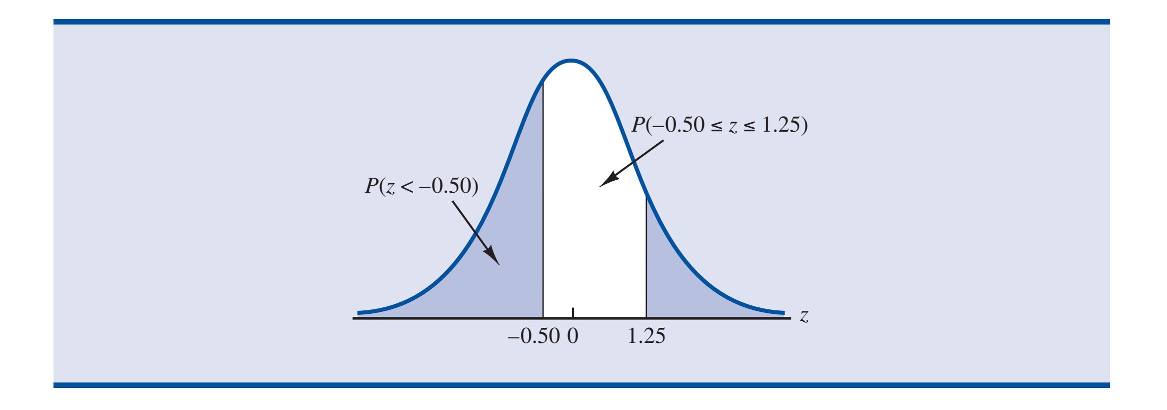


Consulte la página del lado derecho de la tabla de la distribución de probabilidad normal estandar que se encuentra dentro de la cubierta frontal del libro. Esta probabilidad acumulada correspondiente a $z = 1.00$ es el valor que en la tabla se localiza en la intersección del renglón cuyo encabezado es 1.0 y la columna cuyo encabezado es 0.00. Primero localice 1.0 en la columna del extremo izquierdo de la tabla y después localice 0.00 en el renglón en la parte superior de la tabla. Observe que en el interior de la tabla, el renglón 1.0 y la columna 0.00 se cruzan en el valor 0.8413; por tanto, $P(z \leq 1.00) = 0.8413$. Estos pasos se muestran en el extracto siguiente de las tablas de probabilidad.

z	**0.00**	**0.01**	**0.02**
.			
.			
.			
0.9	0.8159	0.8186	0.8212
1.0	**0.8413**	0.8438	0.8461
1.1	0.8643	0.8665	0.8686
1.2	0.8849	0.8869	0.8888
.			
.			
.			

$P(z \leq 1.00)$

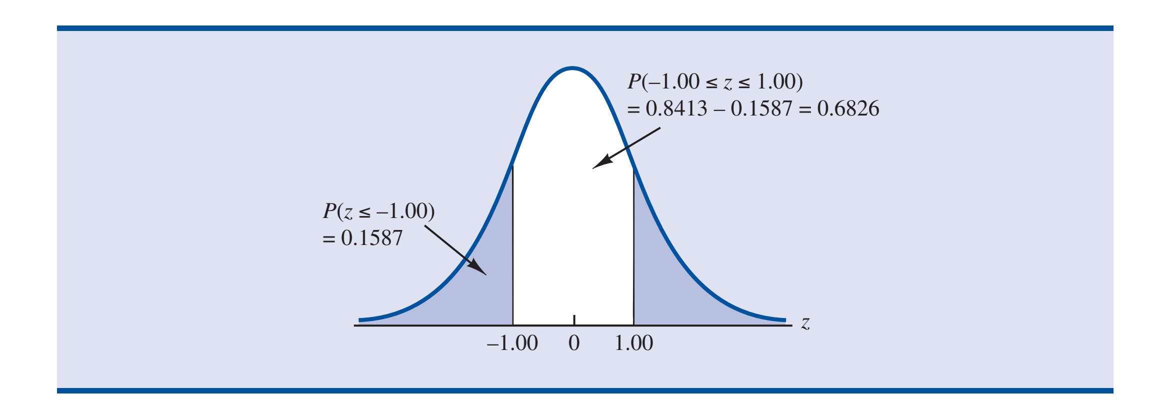
Para ilustrar el segundo tipo de cálculo de una probabilidad se muestra cómo calcular la probabilidad de que z esté en el intervalo entre -0.50 y 1.25; esto es, $P(-.50 \leq z \leq 1.25)$. En la gráfica siguiente se muestra esta área o probabilidad.



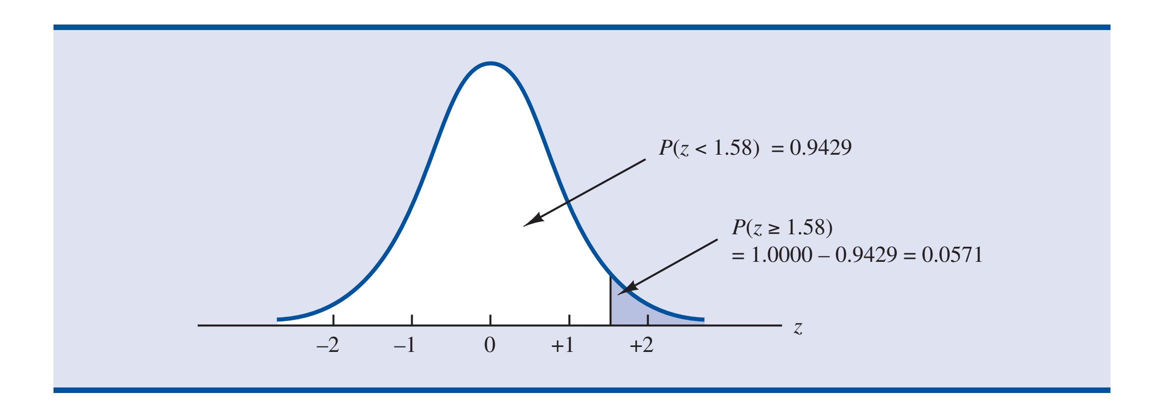
Para calcular esta probabilidad son necesarios tres pasos. Primero, se encuentra el área bajo la curva normal a la izquierda de $z = 1.25$. Segundo, se encuentra el área bajo la curva normal a la izquierda de $z = -0.50$. Por último, se resta el área a la izquierda de $z = -0.50$ del área a la izquierda de $z = 1.25$ y se encuentra, $P(-0.50 \leq z \leq 1.25)$.

Para encontrar el área bajo la curva normal a la izquierda de $z = 1.25$, primero se localiza en la tabla de probabilidad normal estándar el renglón 1.2 y después se avanza por ese renglón hasta la columna 0.05. Como el valor que aparece en el renglón 1.2 columna 0.05 es 0.8944, $P(z \leq 1.25) = 0.8944$. De manera similar, para encontrar el área bajo la curva a la izquierda de $z = -0.50$ se usa el forro izquierdo de la tabla para localizar el valor en el renglón -0.5 columna 0.00; como el valor que se encuentra es 0.3085, $P(z \leq -0.50) = 0.3085$. Por tanto, $P(-0.50 \leq z \leq 1.25) = P(z \leq 1.25) - P(z \leq -0.50) = 0.8944 - 0.3085 = 0.5859$.

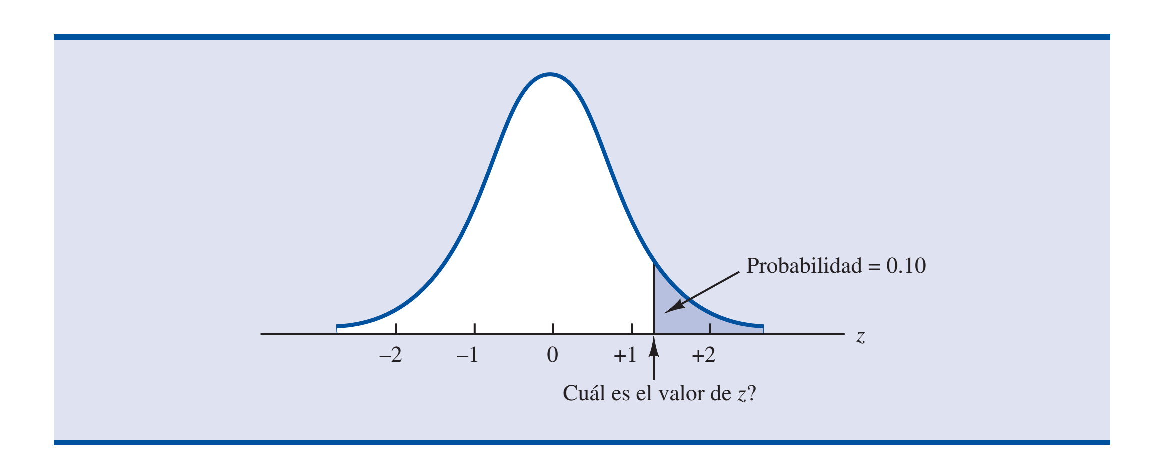
A continuación se presenta otro ejemplo para calcular la probabilidad de que z esté en el intervalo entre dos valores dados. Con frecuencia se desea calcular la probabilidad de que una variable aleatoria normal tome un valor dentro de cierto número de desviaciones estándar respecto a la media. Suponga que desea calcular la probabilidad de que la variable aleatoria normal estándar se encuentre a no más de una desviación estándar de la media; es decir, $P(-1.00 \leq z \leq 1.00)$. Para calcular esta probabilidad tiene que hallar el área bajo la curva entre -1.00 y 1.00. Antes encontró que $P(z \leq 1.00) = 0.8413$. Si va al forro izquierdo, encontrará que el área bajo la curva a la izquierda de $z = -1.00$ es 0.1587, de manera que $P(z \leq -1.00) = 0.1587$. Por tanto, $P(-1.00 \leq z \leq 1.00) = P(z \leq 1.00) - P(z \leq -1.00) = 0.8413 - 0.1587 = 0.6826$. Esta probabilidad se muestra en forma gráfica en la figura siguiente.



Para ilustrar cómo se calcula el tercer tipo de probabilidad, suponga que desea calcular la probabilidad de tener un valor z por lo menos igual a 1.58; es decir, $P(z \geq 1.58)$. El valor en el renglón $z = 1.5$, columna 0.08 de la tabla normal acumulada es 0.9429; por tanto, $P(z < 1.58) = 0.9429$. Pero, como toda el área bajo la curva normal es 1, $P(z \geq 1.58) = 1 - 0.9429 = 0.0571$. En la figura siguiente se muestra esta probabilidad.



En los ejemplos anteriores se muestra cómo calcular probabilidades dados determinados valores de z. En algunas situaciones se da una probabilidad y se trata de hacer lo contrario, encontrar el correspondiente valor de z. Suponga que desea hallar un valor z tal que la probabilidad de obtener un valor z mayor sea 0.10. En la figura siguiente se muestra en forma gráfica esta situación.



Dada una probabilidad, se puede usar la tabla normal estándar para encontrar el correspondiente valor de z.

Este problema es la situación contraria a la presentada en los ejemplos anteriores, en ellos se dio el valor z y se halló la probabilidad o área correspondiente. En este ejemplo se da la probabilidad, o el área, y se pide hallar el valor correspondiente de z. Para esto se usa la tabla de probabilidad normal estándar de una manera un poco diferente.

Recuerde que la tabla de probabilidad normal estándar da el área bajo la curva a la izquierda de un determinado valor z. Se ha recibido la información de que el área en la cola superior (derecha) de la curva es 0.10. Por tanto, el área bajo la curva a la izquierda del valor desconocido de z debe ser 0.9000. Al recorrer el cuerpo de la tabla, se encuentra que 0.8997 es la probabilidad acumulada más cercana a 0.9000. A continuación se reproduce la sección de la tabla en la que se encuentra este resultado.

z	**0.06**	**0.07**	**0.08**	**0.09**
.				
.				
.				
1.0	0.8554	0.8577	0.8599	0.8621
1.1	0.8770	0.8790	0.8810	0.8830
1.2	0.8962	0.8980	0.8997	0.9015
1.3	0.9131	0.9147	0.9162	0.9177
1.4	0.9279	0.9292	0.9306	0.9319
.				
.				
.				

Probabilidad acumulada más cercana a 0.9000

Al leer el valor de z en la columna del extremo izquierdo y en el renglón superior de la tabla, se encuentra que el valor de z es 1.28. De manera que un área de aproximadamente 0.9000 (en realidad de 0.8997) es la que se encuentra a la izquierda de $z = 1.28$.* En términos de la pregunta originalmente planteada, 0.10 es la probabilidad aproximada de que z sea mayor que 1.28.

Estos ejemplos ilustran que la tabla de probabilidades acumuladas para la distribución de probabilidad normal estándar se puede usar para hallar probabilidades correspondientes a valores de la variable aleatoria normal estándar z. Es posible hacer dos tipos de preguntas. En el primer tipo de pregunta se dan valores, o un valor de z, y se pide usar la tabla para determinar el área o probabilidad correspondiente. En el segundo tipo de pregunta se da un área, o probabilidad, y se pide usar la tabla para encontrar el correspondiente valor de z. Por tanto, se necesita tener flexibilidad para usar la tabla de probabilidad normal estándar para responder la pregunta deseada. En la mayoría de los casos, hacer un bosquejo de la gráfica de la distribución de probabilidad normal estándar y sombrear el área deseada será una ayuda para visualizar la situación y encontrar la respuesta correcta.

Cálculo de probabilidades en cualquier distribución de probabilidad normal

La razón por la cual la distribución normal estándar se ha visto de manera tan amplia es que todas las distribuciones normales son calculadas mediante la distribución normal estándar. Esto es, cuando distribución normal con una media μ cualquiera y una desviación estándar σ cualquiera, las preguntas sobre las probabilidades en esta distribución se responden pasando primero a la distribución normal estándar. Use las tablas de probabilidad normal estándar y los valores apropiados de z para hallar las probabilidades deseadas. A continuación se da la fórmula que se emplea para convertir cualquier variable aleatoria x con media μ y desviación estándar σ en la variable aleatoria normal estándar z.

La fórmula para la variable aleatoria normal estándar es semejante a la fórmula que se dio en el capítulo 3 para los puntos z de un conjunto de datos.

CONVERSIÓN A LA VARIABLE ALEATORIA NORMAL ESTÁNDAR

$$z = \frac{x - \mu}{\sigma} \tag{6.3}$$

* Se podía haber hecho una interpolación en el cuerpo de la tabla para obtener una aproximación más exacta al valor z que corresponde al área 0.9000. Al hacerlo en busca de un lugar decimal más preciso se obtiene 1.282. Sin embargo, en la mayor parte de las situaciones prácticas, es suficiente con la precisión obtenida usando el valor más cercano al valor deseado que da la tabla.

Un valor x igual a su media μ da como resultado $z = (\mu - \mu)/\sigma = 0$. De manera que un valor x igual a su media corresponde a $z = 0$. Ahora suponga que x se encuentra una desviación estándar arriba de su media. Es decir, $x = \mu + \sigma$. Aplicando la ecuación (6.3) el valor correspondiente es $z = [(\mu + \sigma) - \mu]/\sigma = \sigma/\sigma = 1$. Así que un valor de x que es una desviación estándar mayor que su media corresponde a $z = 1$. En otras palabras, *z se interpreta como el número de desviaciones estándar a las que está una variable aleatoria* x *de su media μ.*

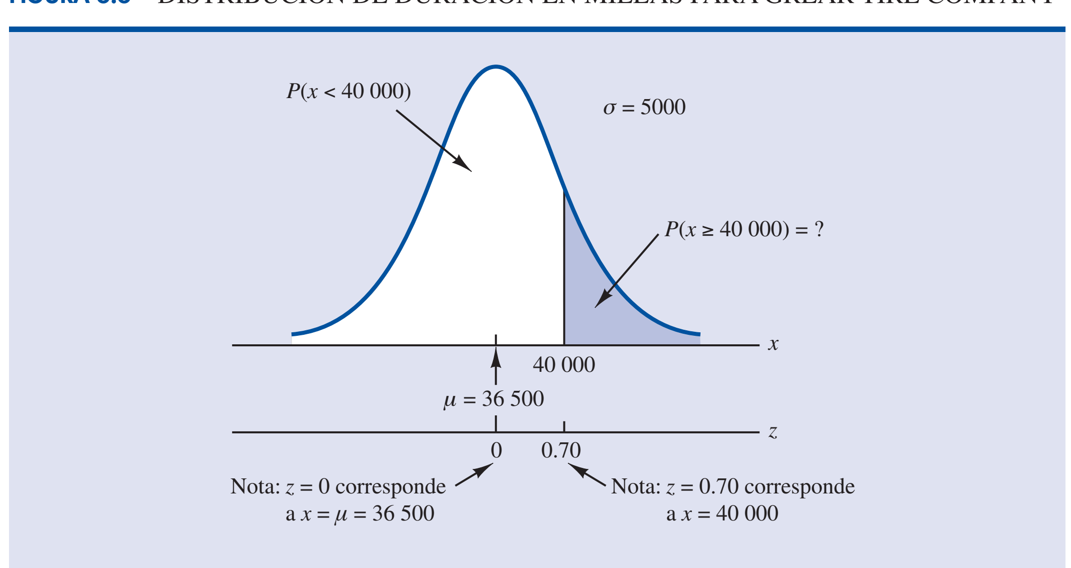
Para ver cómo esta distribución permite calcular probabilidades en cualquier distribución normal, admita que tiene una distribución en la que $\mu = 10$ y $\sigma = 2$. ¿Cuál es la probabilidad de que la variable aleatoria x esté entre 10 y 14? Empleando la ecuación (6.3) se ve que para $z = (x - \mu)/\sigma = (10 - 10)/2 = 0$ y que para $x = 14$, $z = (14 - 10)/2 = 4/2 = 2$. Así, la respuesta a la pregunta acerca de la probabilidad de que x esté entre 10 y 14 está dada por la probabilidad equivalente de que z esté entre 0 y 2 en la distribución normal estándar. En otras palabras, la probabilidad que se está buscando es que la variable aleatoria x esté entre su media y dos desviaciones estándar arriba de la media. Usando $z = 2$ y la tabla de probabilidad normal estándar del forro interior, se ve que $P(z \leq 2) = 0.9772$. Como $P(z \leq 0) = 0.5000$, se tiene que $P(0.00 \leq z \leq 2.00) = P(z \leq 2) - P(z \leq 0) = 0.9772 - 0.5000 = 0.4772$. Por tanto, la probabilidad de que x esté entre 10 y 14 es 0.4772.

El problema de la empresa Grear Tire

Para una aplicación de la distribución de probabilidad normal, suponga que Grear Tire Company ha fabricado un nuevo neumático que será vendido por una cadena nacional de tiendas de descuento. Como este neumático es un producto nuevo, los directivos de Grear piensan que la garantía de duración será un factor importante en la aceptación del neumático. Antes de finalizar la póliza de garantía, los directivos necesitan información probabilística acerca de x = duración del neumático en número de millas.

De acuerdo con las pruebas realizadas al neumático, los ingenieros de Grear estiman que la duración media en millas es $\mu = 36\,500$ millas y que la desviación estándar es $\sigma = 5000$. Además, los datos recogidos indican que es razonable suponer una distribución normal. ¿Qué porcentaje de los neumáticos se espera que duren más de 40 000 millas? En otras palabras, ¿cuál es la probabilidad de que la duración de los neumáticos sea superior a 40 000? Esta pregunta se responde hallando el área de la región sombreada que se observa en la gráfica de la figura 6.6.

FIGURA 6.6 DISTRIBUCIÓN DE DURACIÓN EN MILLAS PARA GREAR TIRE COMPANY



Para $x = 40\,000$, se tiene

$$z = \frac{x - \mu}{\sigma} = \frac{40\,000 - 36\,500}{5000} = \frac{3500}{5000} = 0.70$$

Observe que en la parte inferior de la figura 6.6 el valor $x = 40\,000$ en la distribución normal de Grear Tire corresponde a $z = 0.70$ en la distribución normal estándar. Mediante la tabla de probabilidad normal estándar se encuentra que el área bajo la curva normal estándar a la izquierda de $z = 0.70$ es 0.7580. De manera que $1.000 - 0.7580 = 0.2420$ es la probabilidad de que z sea mayor a 0.70 y por tanto de que x sea mayor a 40 000. Entonces 24.2% de los neumáticos durará más de 40 000 millas.

Ahora suponga que Grear está considerando una garantía que dé un descuento en la sustitución del neumático original si éste no dura lo que asegura la garantía. ¿Cuál deberá ser la duración en millas especificada en la garantía si Grear desea que no más de 10% de los neumáticos alcancen la garantía? En la figura 6.7 se plantea esta pregunta en forma gráfica.

De acuerdo con la figura 6.7, el área bajo la curva a la izquierda de la cantidad desconocida de millas para la garantía debe ser 0.10. De manera que primero se debe encontrar el valor de z que deja un área de 0.10 en el extremo de la cola izquierda de la distribución normal estándar. Según la tabla de probabilidad normal estándar $z = -1.28$ deja un área de 0.10 en el extremo de la cola izquierda. Por tanto, $z = -1.28$ es el valor de la variable aleatoria normal estándar que corresponde a las millas de duración deseadas para la garantía en la distribución normal de Grear Tire. Para hallar el valor de x que corresponde a $z = -1.28$, se tiene

Las millas de garantía que se desean encontrar están a 1.28 desviaciones estándar abajo de la media. Por tanto, $x = \mu - 1.28\sigma$.

$$z = \frac{x - \mu}{\sigma} = -1.28$$

$$x - \mu = -1.28\sigma$$

$$x = \mu - 1.28\sigma$$

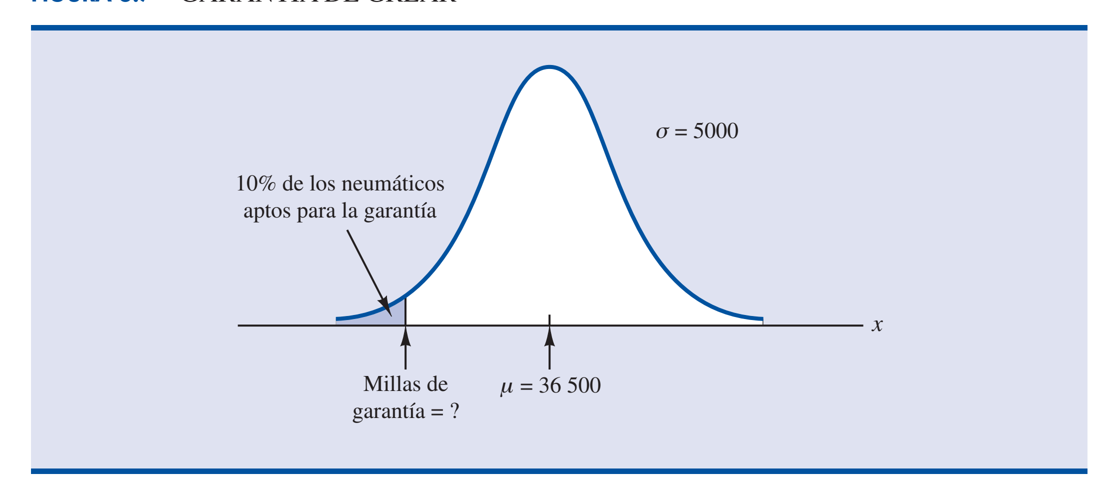
Como $\mu = 36\,500$ y $s = 5000$,

$$x = 36\,500 - 1.28(5000) = 30\,100$$

Al establecer una garantía a partir de 30 000 millas, el porcentaje real apto para la garantía será 9.68%.

Por tanto, una garantía de 30 100 millas cumplirá con el requerimiento de que aproximadamente 10% de los neumáticos sean aptos para la garantía. Con esta información, quizá la empresa establezca una garantía de 30 000 millas.

FIGURA 6.7 GARANTÍA DE GREAR



Nuevamente, se observa la importancia de las distribuciones de probabilidad en el suministro de información para la toma de decisiones. A saber, una vez que la distribución de probabilidad es establecida para una aplicación en particular, puede ser usada para obtener información probabilística acerca del problema. La probabilidad no recomienda directamente una decisión, pero suministra información que ayuda a tomarla entendiendo mejor los riesgos y la incertidumbre asociados al problema. Finalmente, esta información ayuda a enriquecer una buena decisión.

Ejercicios

Métodos

8. Usando como guía la figura 6.4, dibuje la curva normal de la variable aleatoria x cuya media es $\mu = 100$ con desviación estándar de $\sigma = 10$. Indique en el eje horizontal los valores 70, 80, 90, 100, 110, 120 y 130.
9. Una variable aleatoria es normalmente distribuida con media $\mu = 50$ y desviación estándar $\sigma = 5$.
   a. Dibuje la curva normal de la función de densidad de probabilidad. En el eje horizontal dé los valores 35, 40, 45, 50, 55, 60 y 65. En la figura 6.4 se observa que la curva normal casi toca el eje horizontal en los puntos que se encuentran tres desviaciones estándar arriba de la media y tres desviaciones estándar debajo de la media (en este caso en 35 y 65).
   b. ¿Cuál es la probabilidad de que la variable aleatoria tome un valor entre 45 y 55?
   c. ¿De que la variable aleatoria tome un valor entre 40 y 60?
10. Dibuje la gráfica de la distribución normal estándar. Etiquete el eje horizontal con los valores -3, -2, -1, 0, 1, 2 y 3. Después use la tabla de probabilidades de la distribución normal estándar que se encuentra en el forro interior del libro para calcular las probabilidades siguientes.
    a. $P(z \leq 1.5)$
    b. $P(z \leq 1)$
    c. $P(1 \leq z \leq 1.5)$
    d. $P(0 < z < 2.5)$
11. Dado que z es la variable normal estándar, calcule las probabilidades siguientes.
    a. $P(z \leq -1.0)$
    b. $P(z \geq -1)$
    c. $P(z \geq -1.5)$
    d. $P(-2.5 \leq z)$
    e. $P(-3 < z \leq 0)$
12. Dado que z es la variable normal estándar, calcule las probabilidades siguientes.
    a. $P(0 \leq z \leq 0.83)$
    b. $P(-1.57 \leq z \leq 0)$
    c. $P(z > 0.44)$
    d. $P(z \geq -0.23)$
    e. $P(z < 1.20)$
    f. $P(z \leq -0.71)$



13. Dado que z es la variable normal estándar, calcule las probabilidades siguientes.
    a. $P(-1.98 \leq z \leq 0.49)$
    b. $P(0.52 \leq z \leq 1.22)$
    c. $P(-1.75 \leq z \leq -1.04)$
14. Dado que z es la variable normal estándar, encuentre z en cada una de las situaciones siguientes.
    a. El área a la izquierda de z es 0.9750.
    b. El área entre 0 y z es 0.4750.
    c. El área a la izquierda de z es 0.7291.
    d. El área a la derecha de z es 0.1314.
    e. El área a la izquierda de z es 0.6700.
    f. El área a la derecha de z es 0.3300.



15. Dado que z es la variable normal estándar, halle z en cada una de las situaciones siguientes.
    a. El área a la izquierda de z es 0.2119
    b. El área entre $-z$ y z es 0.9030.
    c. El área entre $-z$ y z es 0.2052.
    d. El área a la izquierda de z es 0.9948.
    e. El área a la derecha de z es 0.6915.
16. Dado que z es la variable normal estándar, encuentre z en cada una de las situaciones siguientes.
    a. El área a la derecha de z es 0.01
    b. El área a la derecha de z es 0.025.
    c. El área a la derecha de z es 0.05.
    d. El área a la derecha de z es 0.10.

Aplicaciones

17. Una persona con una buena historia crediticia tiene una deuda promedio de $15 015 (*BusinessWeek*, 20 de marzo de 2006). Suponga que la desviación estándar es de $3 540 y que los montos de las deudas están distribuidos normalmente.
    a. ¿Cuál es la probabilidad de que la deuda de una persona con buena historia crediticia sea mayor a $18 000?
    b. ¿De que la deuda de una persona con buena historia crediticia sea de menos de $10 000?
    c. ¿De que la deuda de una persona con buena historia crediticia esté entre $12 000 y $18 000?
    d. ¿De que la deuda de una persona con buena historia crediticia sea mayor a $14 000?



18. El precio promedio de las acciones que pertenecen a S&P500 es de $30 y la desviación estándar es $8.20 (*BusinessWeek*, Special Annual Issue, primavera de 2003). Suponga que los precios de las acciones están distribuidos normalmente.
    a. ¿Cuál es la probabilidad de que el precio de las acciones de una empresa sea por lo menos de $40?
    b. ¿De que el precio de las acciones de una empresa no sea mayor a $20?
    c. ¿De cuánto deben ser los precios de las acciones de una empresa para que esté entre las 10% mejores?
19. La cantidad promedio de precipitación pluvial en Dallas, Texas, durante el mes de abril es 3.5 pulgadas (*The World Almanac*, 2000). Suponga que se puede usar una distribución normal y que la desviación estándar es 0.8 pulgadas.
    a. ¿Qué porcentaje del tiempo la precipitación pluvial en abril es mayor que 5 pulgadas?
    b. ¿Qué porcentaje del tiempo la precipitación pluvial en abril es menor que 3 pulgadas?
    c. Un mes se considera como extremadamente húmedo si la precipitación pluvial es 10% superior para ese mes. ¿Cuánta debe ser la precipitación pluvial en abril para que sea considerado un mes extremadamente húmedo?
20. En enero de 2003 un empleado estadounidense pasaba, en promedio, 77 horas conectado a Internet durante las horas de trabajo (CNBC, 15 de marzo de 2003). Suponga que la media poblacional es 77 horas, tiempos que están distribuidos normalmente y que la desviación estándar es 20 horas.
    a. ¿Cuál es la probabilidad de que en enero de 2003 un empleado seleccionado aleatoriamente haya pasado menos de 50 horas conectado a Internet?
    b. ¿Qué porcentaje de los empleados pasó en enero de 2003 más de 100 horas conectado a Internet?
    c. Un usuario es clasificado como intensivo si se encuentra en el 20% superior de uso. ¿Cuántas horas tiene un empleado que haber estado conectado a Internet en enero de 2003 para que se le considerara un usuario intensivo?
21. La puntuación de una persona en una prueba de IQ debe estar en el 2% superior para que sea clasificado como miembro del grupo Mensa, la sociedad internacional de IQ elevado (*U.S. Airways Attaché*, septiembre de 2000). Si las puntuaciones de IQ tienen una distribución normal con una media de 100 y desviación estándar de 15, ¿cuál debe ser la puntuación de una persona para que se le considere miembro del grupo Mensa?

22. La tasa de remuneración media por hora para administrativos financieros en una determinada región es \$32.62 y la desviación estándar es \$2.32 (Bureau of Labor Statistics, septiembre de 2005). Suponga que estas tasas de remuneración están distribuidas normalmente.
    a. ¿Cuál es la probabilidad de que un directivo financiero tenga una remuneración entre \$30 y \$35 por hora?
    b. ¿Qué tan alta debe ser la remuneración por hora para que un directivo financiero tenga un pago 10% superior?
    c. ¿Cuál es la probabilidad de que la remuneración por hora de un directivo financiero sea menos de \$28 por hora?
23. El tiempo necesario para hacer un examen final en un determinado curso de una universidad tiene una distribución normal cuya media es 80 minutos con desviación estándar de 10 minutos. Conteste las preguntas siguientes
    a. ¿Cuál es la probabilidad de terminar el examen en una hora o menos?
    b. ¿Cuál es la probabilidad de que un estudiante termine el examen en más de 60 minutos pero en menos de 75 minutos?
    c. Suponga que en la clase hay 60 estudiantes y que el tiempo para resolver el examen es de 90 minutos. ¿Cuántos estudiantes piensa usted que no podrán terminar el examen en este tiempo?



Volume

24. El volumen de negociaciones en la Bolsa de Nueva York es más intenso en la primera media hora (en la mañana temprano) y la última media hora (al final de la tarde) de un día de trabajo. A continuación se presentan los volúmenes (en millones de acciones) de 13 días de enero y febrero.

214	163	265	194	180
202	198	212	201	
174	171	211	211	

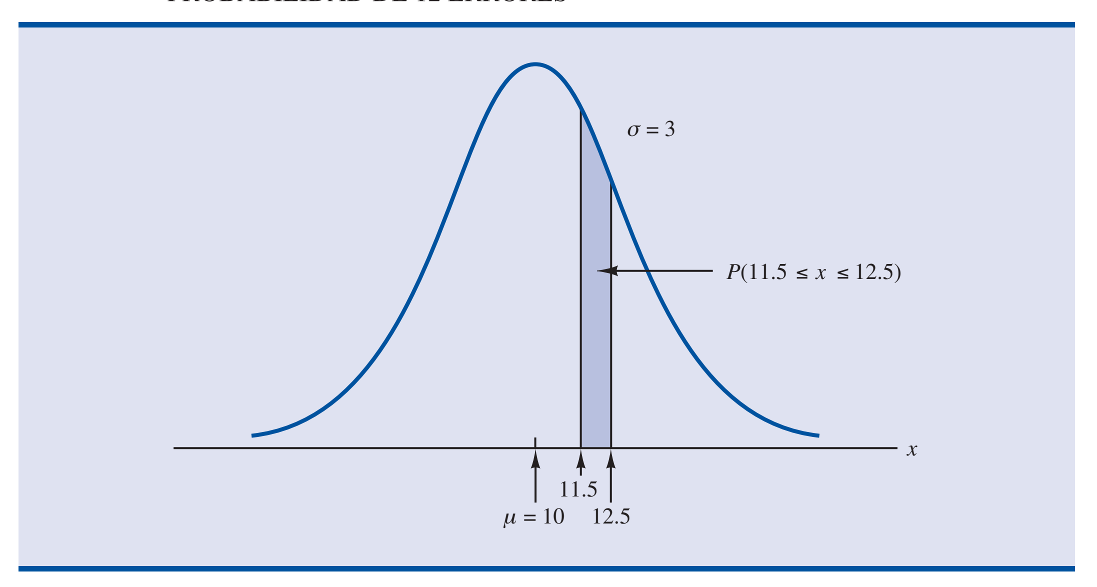
La distribución de probabilidad de los volúmenes de negociaciones es aproximadamente normal.
    a. Calcule la media y la desviación estándar a usar como estimaciones de la media y de la desviación estándar de la población.
    b. ¿Cuál es la probabilidad de que, en un día elegido al azar, el volumen de negociaciones en la mañana temprano sea superior a 180 millones de acciones?
    c. ¿Cuál es la probabilidad de que, en un día elegido al azar, el volumen de negociaciones en la mañana temprano sea superior a 230 millones de acciones?
    d. ¿Cuántas acciones deberán ser negociadas para que el volumen de negociaciones en la mañana temprano de un día determinado pertenezca al 5% de los días de mayor movimiento?
25. De acuerdo con la Sleep Foundation, en promedio se duermen 6 .8 horas por noche. Suponga que la desviación estándar es 0.6 horas y que la distribución de probabilidad es normal.
    a. ¿Cuál es la probabilidad de que una persona seleccionada al azar duerma más de ocho horas?
    b. ¿De que una persona tomada aleatoriamente duerma seis horas o menos?
    c. Los médicos aconsejan dormir entre siete y nueve horas por noche. ¿Qué porcentaje de la población duerme esta cantidad?



6.3 Aproximación normal de las probabilidades binomiales

En la sección 5.4 se presentó la distribución binomial discreta. Recuerde que un experimento binomial consiste en una serie de n ensayos idénticos e independientes, habiendo para cada ensayo dos resultados posibles, éxito o fracaso. La probabilidad de éxito en un ensayo es la misma que en cualquier otro de los ensayos y se denota p. La variable aleatoria binomial es el número de éxitos en n ensayos y lo que se quiere saber es la probabilidad de x éxitos en n ensayos.

FIGURA 6.8 APROXIMACIÓN NORMAL A UNA PROBABILIDAD BINOMIAL DISTRIBUCIÓN EN LA QUE $n = 100$ Y $p = 0.10$ MOSTRANDO LA PROBABILIDAD DE 12 ERRORES



La evaluación de una función de probabilidad binomial, a mano o con una calculadora, se dificulta cuando el número de ensayos es muy grande. En los casos en que $np \geq 5$ y $n(1 - p) \geq 5$, la distribución normal proporciona una aproximación a las probabilidades binomiales que es fácil de usar. Cuando se usa la aproximación normal a la binomial, en la definición de la curva normal $\mu = np$ y $\sigma = \sqrt{np(1 - p)}$.

Para ilustrar la aproximación normal a la binomial, suponga que una empresa sabe por experiencia que 10% de sus facturas tienen algún error. Toma una muestra de 100 facturas y desea calcular la probabilidad de que 12 de estas facturas contengan algún error. Es decir, quiere hallar la probabilidad binomial de 12 éxitos en 100 ensayos. Aplicando la aproximación normal a este caso se tiene, $\mu = np = (100)(0.1) = 10$ y $\sigma = \sqrt{np(1 - p)} = \sqrt{(100)(0.1)(0.9)} = 3$. En la figura 6.8 se muestra la distribución normal con $\mu = 10$ y $\sigma = 3$.

Recuerde que en una distribución de probabilidad continua las probabilidades se calculan como áreas bajo la curva de la función de densidad de probabilidad. En consecuencia, la probabilidad que tiene un solo valor de la variable aleatoria es cero. Por tanto, para aproximar la probabilidad binomial de 12 éxitos se calcula el área correspondiente bajo la curva normal entre 11.5 y 12.5. Al 0.5 que se suma y se resta al 12 se le conoce como **factor de corrección por continuidad.** Este factor se introduce debido a que se está usando una distribución continua para aproximar una distribución discreta. Así, $P(x = 12)$ de una distribución binomial *discreta* se aproxima mediante $P(11.5 \leq x \leq 12.5)$ en la distribución normal *continua*.

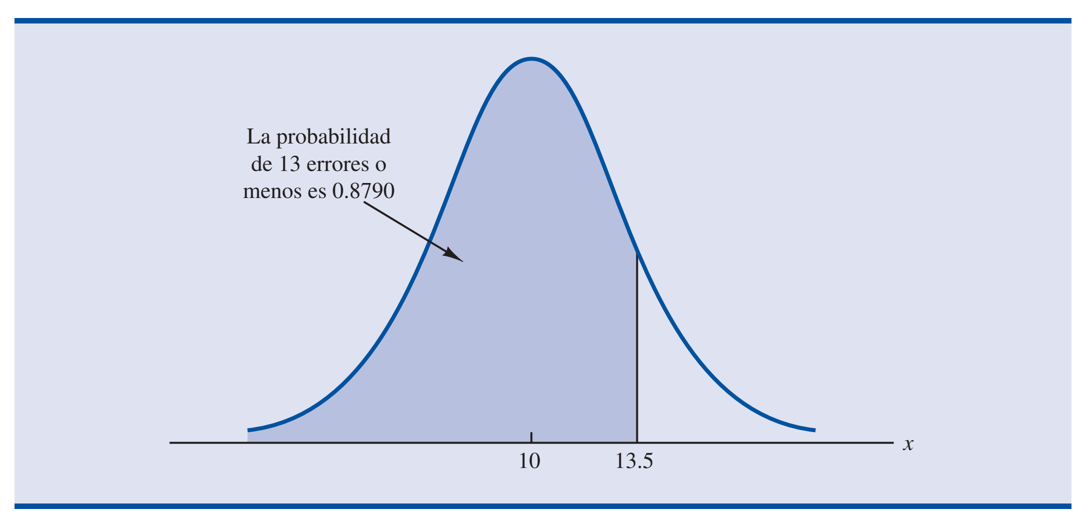
Convirtiendo la distribución normal estándar para calcular $P(11.5 \leq x \leq 12.5)$, se tiene

$$z = \frac{x - \mu}{\sigma} = \frac{12.5 - 10.0}{3} = 0.83 \quad \text{para} \quad x = 12.5$$

y

$$z = \frac{x - \mu}{\sigma} = \frac{11.5 - 10.0}{3} = 0.50 \quad \text{para} \quad x = 11.5$$

FIGURA 6.9 APROXIMACIÓN NORMAL A UNA PROBABILIDAD BINOMIAL: DISTRIBUCIÓN EN LA QUE $n = 100$ Y $p = 0.10$ MUESTRAN LA PROBABILIDAD DE 13 ERRORES O MENOS



En la tabla de la probabilidad normal estándar aparece que el área bajo la curva (figura 6.8) a la izquierda de 12.5 es 0.7967. De manera similar, el área bajo la curva a la izquierda de 11.5 es 0.6915. Por tanto, el área entre 11.5 y 12.5 es $0.7967 - 0.6915 = 0.1052$. El cálculo normal de la probabilidad de 12 éxitos en 100 ensayos es 0.1052.

Para tener un ejemplo más, suponga que se quiere calcular la probabilidad de 13 o menos facturas con errores en una muestra de 100 facturas. En la figura 6.9 se muestra el área bajo la curva que aproxima esta probabilidad. Observe que debido al uso del factor de continuidad el valor que se usa para calcular esta probabilidad es 13.5. El valor z que corresponde a 13.5 es

$$z = \frac{13.5 - 10.0}{3.0} = 1.17$$

En la tabla de probabilidad normal estándar se observa que el área bajo la curva normal estándar y a la izquierda de $z = 1.17$ es 0.8790. El área bajo la curva normal que aproxima la probabilidad de 13 o menos facturas con errores es la porción sombreada que se observa en la gráfica de la figura 6.9.

Ejercicios

Métodos



26. En una distribución de probabilidad binomial con $p = 0.20$ y $n = 100$.
    a. ¿Cuál es la media y la desviación estándar?
    b. ¿En esta situación las probabilidades binomiales pueden ser aproximadas por la distribución de probabilidad normal? Explique.
    c. ¿Cuál es la probabilidad de exactamente 24 éxitos?
    d. ¿Cuál es la probabilidad de 18 a 22 éxitos?
    e. ¿Cuál es la probabilidad de 15 o menos éxitos?
27. Suponga que se tiene una distribución de probabilidad binomial en la que $p = 0.60$ y $n = 200$.
    a. ¿Cuál es la media y la desviación estándar?
    b. ¿En esta situación las probabilidades binomiales puedan ser aproximadas por la distribución de probabilidad normal? Explique.
    c. ¿Cuál es la probabilidad de 100 a 110 éxitos?

d. ¿Cuál es la probabilidad de 130 o más éxitos?
e. ¿Cuál es la ventaja de usar la distribución de probabilidad normal para aproximar las probabilidades binomiales? Use el inciso d para explicar las ventajas.

Aplicaciones



28. El presidente Bush propuso eliminar los impuestos sobre los dividendos que pagan los accionistas debido a que esto resulta en un doble pago de impuestos. Las ganancias que se usan para pagar los dividendos ya han sido grabadas. En un sondeo sobre este tema se encontró que 47% de los estadounidenses estaban a favor de esta propuesta. La posición de los partidos políticos era 64% de los republicanos y 29% de los demócratas a favor de la propuesta (*Investor´s Business Daily*, 13 de enero de 2003). Suponga que 250 estadounidenses se reúnen para una conferencia acerca de la propuesta.
    a. ¿Cuál es la probabilidad de que por lo menos la mitad del grupo esté a favor de la propuesta?
    b. Más tarde se entera de que en el grupo hay 150 republicanos y 100 demócratas. Ahora, ¿cuál es su estimación del número esperado a favor de la propuesta?
    c. Ahora que conoce la composición del grupo, ¿cree que un conferencista a favor de la propuesta sea mejor recibido que uno que esté en contra de la propuesta?
29. La tasa de desempleo es de 5.8% (Bureau of Labor Statistics, www.bls.gov, 3 de abril de 2003). Suponga que se seleccionan aleatoriamente 100 personas que se pueden emplear.
    a. ¿Cuál es el número esperado de quienes están desempleados?
    b. ¿Cuál es la varianza y la desviación estándar del número de los que están desempleados?
    c. ¿Cuál es la probabilidad de que exactamente seis estén desempleados?
    d. ¿Cuál es la probabilidad de que por lo menos cuatro estén desempleados?
30. Cuando usted firma un contrato para una tarjeta de crédito, ¿lee cuidadosamente el contrato? En un sondeo FindLaw.com le preguntó a las personas "¿Qué tan cuidadosamente lee usted un contrato para una tarjeta de crédito?" Los hallazgos fueron que 44% leen cada palabra, 33% leen lo suficiente para entender el contrato, 11% sólo le echa una mirada y 4% no lo leen en absoluto.
    a. En una muestra de 500 personas ¿cuántas esperaría usted que respondan que leen cada palabra de un contrato para una tarjeta de crédito?
    b. En una muestra de 500 personas ¿cuál es la probabilidad de que 200 o menos digan que leen cada palabra de un contrato para una tarjeta de crédito?
    c. En una muestra de 500 personas ¿cuál es la probabilidad de que por lo menos 15 digan que no leen en absoluto un contrato para una tarjeta de crédito?
31. El Myrtle Beach hotel tiene 120 habitaciones. En los meses de primavera su ocupación es de 75%.
    a. ¿Cuál es la probabilidad de que por lo menos la mitad de las habitaciones estén ocupadas en un día dado?
    b. ¿De que 100 o más de las habitaciones estén ocupadas en un día dado?
    c. ¿De que 80 o menos de las habitaciones estén ocupadas en un día dado?



Distribución de probabilidad exponencial

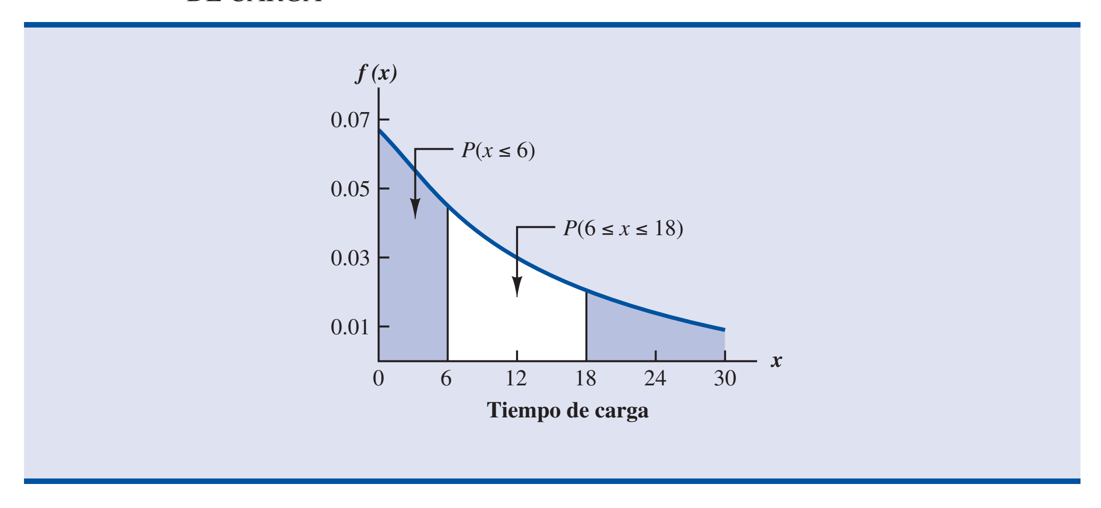
La **distribución de probabilidad exponencial** se aplica a variables como las llegadas de automóviles a un lavado de coches, los tiempos requeridos para cargar un camión, las distancias entre dos averías en una carretera, etc. A continuación se da la función de densidad de probabilidad exponencial.

FUNCIÓN DE DENSIDAD DE PROBABILIDAD EXPONENCIAL

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \qquad \text{para } x \geq 0, \mu > 0 \tag{6.4}$$

donde μ = valor esperado o media

FIGURA 6.10 DISTRIBUCIÓN EXPONENCIAL PARA EL EJEMPLO DEL ÁREA DE CARGA



Como ejemplo de la distribución exponencial, suponga que x representa el tiempo que se necesita para cargar un camión en un área de carga, y que este tiempo de carga sigue una distribución exponencial. Si el tiempo de carga medio o promedio es 15 minutos ($\mu = 15$), la función de densidad de probabilidad apropiada para x es

$$f(x) = \frac{1}{15} e^{-x/15}$$

La figura 6.10 es la gráfica de esta función de densidad de probabilidad.

Cálculo de probabilidades en la distribución exponencial

Como ocurre con cualquier distribución de probabilidad continua, el área bajo la curva correspondiendo a un intervalo da la probabilidad de que la variable aleatoria tome algún valor en ese intervalo. En el ejemplo del área de carga, la probabilidad de que cargar un camión necesite 6 minutos o menos $P\ (x \leq 6)$ está definida como el área bajo la curva de la figura 10.6 que va desde $x = 0$ hasta $x = 6$. De manera similar, la probabilidad de que el tiempo de carga sean 18 minutos o menos $P\ (x \leq 18)$ es el área bajo la curva desde $x = 0$ hasta $x = 18$. Observe también que la probabilidad de que el tiempo de carga esté entre 6 y 18 minutos $P\ (6 \leq x \leq 18)$ corresponde al área bajo la curva desde $x = 6$ hasta $x = 18$.

En aplicaciones de colas de espera, la distribución exponencial suele emplearse para tiempos de servicio.

Para calcular probabilidades exponenciales como las que se acaban de describir, se usa la fórmula siguiente. Esta fórmula aporta la probabilidad acumulada de obtener un valor de la variable aleatoria exponencial que sea menor o igual que algún valor específico denotado por x_0.

DISTRIBUCIÓN EXPONENCIAL: PROBABILIDADES ACUMULADAS

$$P(x \leq x_0) = 1 - e^{-x_0/\mu} \quad \textbf{(6.5)}$$

En el ejemplo del área de carga, x = tiempo de carga en minutos y $\mu = 15$ minutos. A partir de la ecuación (6.5)

$$P(x \leq x_0) = 1 - e^{-x_0/15}$$

Por tanto, la probabilidad de que cargar un camión requiera 6 minutos o menos es

$$P(x \leq 6) = 1 - e^{-6/15} = 0.3297$$

Con la ecuación (6.5) se calcula la probabilidad de que cargar un camión requiera 18 minutos o menos.

$$P(x \leq 18) = 1 - e^{-18/15} = 0.6988$$

De manera que la probabilidad de que para cargar un camión se necesiten entre 6 y 18 minutos es igual a $0.6988 - 0.3297 = 0.3691$. Probabilidades para cualquier otro intervalo se calculan de manera semejante.

La distribución exponencial tiene la propiedad de que la media y la desviación estándar son iguales.

En el ejemplo anterior el tiempo medio para cargar un camión fue $\mu = 15$ minutos. La distribución exponencial tiene la propiedad de que la media de la distribución y la desviación estándar de la distribución son iguales. Por tanto, la desviación estándar del tiempo que se necesita para cargar un camión es $\sigma = 15$ minutos y la varianza es $\sigma^2 = (15)^2 = 225$.

Relación entre la distribución de Poisson y la exponencial

En la sección 5.5 se presentó la distribución de probabilidad de Poisson como una distribución de probabilidad discreta que se usa para examinar el número de ocurrencias de un evento en un determinado intervalo de tiempo o de espacio. Recuerde que la función de probabilidad de Poisson es

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

donde

$\mu =$ valor esperado o número medio de ocurrencias en un determinado intervalo

Si las llegadas siguen una distribución de Poisson, el tiempo entre las llegadas debe seguir una distribución exponencial.

La distribución de probabilidad exponencial continua está relacionada con la distribución discreta de Poisson. Si la distribución de Poisson da una descripción del número de ocurrencias por intervalo, la distribución exponencial aporta una descripción de la longitud de los intervalos entre las ocurrencias.

Para ilustrar esta relación, suponga que el número de automóviles que llegan a un lavado de coches durante una hora se describe mediante la distribución de probabilidad de Poisson, con una media de 10 automóviles por hora. La función de probabilidad de Poisson que da la probabilidad de x llegadas por hora es

$$f(x) = \frac{10^x e^{-10}}{x!}$$

Dado que el número promedio de llegadas es 10 automóviles por hora, el tiempo promedio entre las llegadas de los automóviles es

$$\frac{1 \text{ hora}}{10 \text{ automóviles}} = 0.1 \text{ hora/automóvil}$$

Entonces, la distribución exponencial que describe el tiempo entre las llegadas tiene una media de $\mu = 0.1$ por automóvil; la función de densidad de probabilidad exponencial es

$$f(x) = \frac{1}{0.1} e^{-x/0.1} = 10e^{-10x}$$

NOTAS Y COMENTARIOS

Como se observa en la figura 6.10, la distribución exponencial es sesgada a la derecha. En efecto, la medida del sesgo en la distribución exponencial es 2. La distribución exponencial da una idea clara de cómo es una distribución sesgada.

Ejercicios

Métodos

32. Considere la siguiente función de densidad de probabilidad exponencial.

$$f(x) = \frac{1}{8} e^{-x/8} \quad \text{para } x \geq 0$$

a. Halle $P(x \leq 6)$.
b. Encuentre $P(x \leq 4)$.
c. Halle $P(x \geq 6)$.
d. Encuentre $P(4 \leq x \leq 6)$.



33. Considere la siguiente función de densidad de probabilidad exponencial.

$$f(x) = \frac{1}{3} e^{-x/3} \quad \text{para } x \geq 0$$

a. Dé la fórmula para hallar $P(x \leq x_0)$..
b. Halle $P(x \leq 2)$.
c. Encuentre $P(x \geq 3)$.
d. Halle $P(x \leq 5)$.
e. Halle $P(2 \leq x \leq 5)$.

Aplicaciones

34. El tiempo requerido para pasar por la inspección en los aeropuertos puede ser molesto para los pasajeros. El tiempo medio de espera en los periodos pico en el Cincinnnati/Northern Kentucky Internacional Airport es de 12.1 minutos (*The Cincinnati Enquirer*, 2 de febrero de 2006). Suponga que los tiempos para pasar por la inspección de seguridad tienen una distribución exponencial.
   a. ¿Cuál es la probabilidad de que durante los periodos pico se requieran 10 minutos para pasar la inspección de seguridad?
   b. ¿De que durante los periodos pico se requieran más de 20 minutos para pasar la inspección de seguridad?
   c. ¿De que durante los periodos pico se requieran entre 10 y 20 minutos para pasar la inspección de seguridad?
   d. Son las 8 de la mañana (periodo pico) y usted se acaba de formar en la fila para la inspección de seguridad. Para alcanzar su avión, tiene que estar en la puerta de arribo en no más de 30 minutos. Si necesitara 12 minutos una vez pasada la inspección de seguridad para llegar a la puerta de arribo, ¿cuál es la probabilidad de que pierda el avión?



35. Los tiempos entre las llegadas de vehículos a un determinado entronque siguen una distribución de probabilidad exponencial cuya media es 12 segundos.
   a. Dibuje esta distribución de probabilidad exponencial.
   b. ¿Cuál es la probabilidad de que los tiempos de llegada entre vehículos sean 12 segundos o menos?

c. ¿Cuál es la probabilidad de que los tiempos de llegada entre vehículos sean 6 segundos o menos?
d. ¿Cuál es la probabilidad de 30 o más segundos entre los tiempos de llegada?

36. El tiempo de vida (en hora) de un dispositivo electrónico es una variable aleatoria que tiene la siguiente función de densidad de probabilidad exponencial.

$$f(x) = \frac{1}{50} e^{-x/50} \qquad \text{para } x \geq 0$$

a. ¿Cuál es la media del tiempo de vida de este dispositivo?
b. ¿Cuál es la probabilidad de que el dispositivo tenga una falla en las primeras 25 horas de funcionamiento?
c. ¿Cuál es la probabilidad de que el dispositivo funcione 100 o más horas sin fallar?

37. Sparagowsky & Associates hace un estudio sobre los tiempos necesarios para atender a un cliente en la ventanilla de su automóvil en los restaurantes de comida rápida. En McDonald´s el tiempo medio para atender a un cliente fue 2.78 minutos (*The Cincinnati Enquirer*, 9 de julio de 2000). Tiempos de servicio como los de estos restaurantes suelen seguir una distribución exponencial.
a. ¿Cuál es la probabilidad de que el tiempo para atender a un cliente sea menor que 2 minutos?
b. ¿De que el tiempo para atender a un cliente sean más de 5 minutos?
c. ¿De que el tiempo para atender a un cliente sean más de 2.78 minutos?

38. ¿Las interrupciones durante su trabajo reducen su productividad? De acuerdo con un estudio realizado por la University of California–Irvine, las personas de negocios son interrumpidas aproximadamente 51/2 veces por hora (*Fortune*, 20 de marzo de 2006). Suponga que el número de interrupciones sigue una distribución de probabilidad de Poisson.
a. Dé la distribución de probabilidad para el tiempo entre las interrupciones.
b. ¿Cuál es la probabilidad de que una persona de negocios no tenga ninguna interrupción en 15 minutos?
c. ¿Cuál es la probabilidad de que la siguiente interrupción a una determinada persona de negocios ocurra en no más de 10 minutos?

Resumen

En este capítulo se amplía el estudio de las distribuciones de probabilidad al caso de las variables aleatorias continuas. La principal diferencia conceptual entre distribuciones de probabilidades discretas y continuas está en el método para calcular las probabilidades. En el caso de distribuciones discretas la función de probabilidad $f(x)$ da la probabilidad de que la variable aleatoria x tome diversos valores. En el caso de las distribuciones continuas, la función de densidad de probabilidad $f(x)$ no da directamente valores de probabilidad. Aquí, las probabilidades están dadas por áreas bajo la curva o gráfica de la función de densidad de probabilidad $f(x)$. Como el área bajo la curva sobre un solo punto es 0, se observa que en una variable aleatoria continua la probabilidad de cualquier valor particular es 0.

Se vieron a detalle tres distribuciones de probabilidad continua: la uniforme, la normal y la exponencial. La distribución normal es muy empleada en la inferencia estadística y será muy empleada en lo que resta del libro.

Glosario

Función de densidad de probabilidad Función que se usa para calcular probabilidades de una variable aleatoria continua. El área bajo la gráfica de una función de densidad de probabilidad y sobre un intervalo representa probabilidad.

Distribución de probabilidad uniforme Distribución de probabilidad continua en la cual la probabilidad de que una variable aleatoria tome un valor en cualquier intervalo es igual para intervalos de igual longitud.

Distribución de probabilidad normal Una distribución de probabilidad continua. Su función de densidad de probabilidad tiene forma de campana y está determinada por la media μ y la desviación estándar σ.

Distribución de probabilidad normal estándar Distribución normal en la cual la media es cero y la desviación estándar es uno.

Factor de corrección por continuidad Valor de 0.5 que se suma o resta al valor de x cuando se usa una distribución normal continua para aproximar una distribución binomial discreta.

Distribución de probabilidad exponencial Una distribución de probabilidad continua útil para calcular probabilidades acerca del tiempo que se necesita para realizar una tarea.

Fórmulas clave

Función de densidad de probabilidad uniforme

$$f(x) = \begin{cases} \dfrac{1}{b - a} & \text{para } a \leq x \leq b \\ 0 & \text{si no es así} \end{cases} \tag{6.1}$$

Función de densidad de probabilidad normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{6.2}$$

Conversión a la variable aleatoria normal estándar

$$z = \frac{x - \mu}{\sigma} \tag{6.3}$$

Función de densidad de probabilidad exponencial

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \qquad \text{para } x \geq 0, \mu > 0 \tag{6.4}$$

Distribución exponencial: probabilidades acumuladas

$$P(x \leq x_0) = 1 - e^{-x_0/\mu} \tag{6.5}$$

Ejercicios complementarios

39. Una ejecutiva de negocios se va a mudar de Chicago a Atlanta y necesita vender rápidamente su casa en Chicago. Un empleado le ofrece comprársela en $210 000, pero la oferta expira al final de esa semana. En este momento la ejecutiva no tiene otra oferta mejor, pero piensa que puede dejar la casa en el mercado un mes más. De acuerdo con las pláticas que ha tenido con su agente inmobiliario la ejecutiva cree que los precios que pueden ofrecerle dejando la casa un mes más en el mercado están distribuidos uniformemente entre $200 000 y $225 000.
    a. Si deja la casa en el mercado un mes más, ¿cuál es la expresión matemática para la función de densidad de probabilidad de los precios de venta que le sean ofrecidos?
    b. Si la deja en el mercado un mes más, ¿cuál es la probabilidad de que venda la casa en por lo menos $215 000?
    c. Si la deja en el mercado un mes más, ¿cuál es la probabilidad de que venda la casa en menos de $210 000?
    d. ¿Deberá dejar la ejecutiva su casa en el mercado un mes más? ¿Por qué sí o por qué no?

40. La U.S. Bureau of Labor Statistics informa que el gasto promedio anual en alimentos y bebidas de una familia es $5700 (*Money*, diciembre de 2003). Suponga que los gastos anuales en alimentos y bebidas están distribuidos en forma normal y que la desviación estándar es $1500.
    a. ¿Cuál es el intervalo en que se encuentran los gastos de 10% de las familias que tienen los menores gastos anuales en alimentos y bebidas?
    b. ¿Qué porcentaje de las familias gasta más de $7000 anualmente en alimentos y bebidas?
    c. ¿Cuál es el intervalo en el que se encuentran los gastos de 5% de las familias que tienen los gastos más altos en alimentos y bebidas?
41. Motorola usa la distribución normal para determinar la probabilidad de defectos y el número de defectos esperados en un proceso de producción. Suponga que en un proceso de producción el peso promedio de los artículos producidos es 10 onzas. Calcule la probabilidad de un defecto y el número esperado de defectos en 100 unidades producidas en las situaciones siguientes.
    a. La desviación estándar del proceso es 0.15 y los límites para el proceso se han fijado en más o menos una desviación estándar. Las unidades que pesen más de 9.85 o menos de 10.15 onzas se clasifican como defectuosas.
    b. Después de hacer mejoras al proceso de producción, la desviación estándar se reduce a 0.05. Asuma que se siguen usando los mismos límites para el proceso; artículos que pesen menos de 9.85 o más de 10.15 onzas se clasifican como defectuosos.
    c. ¿Cuál es la ventaja de haber reducido la variación en el proceso de producción, haciendo que los límites se encuentren a un número mayor de desviaciones estándar de la media?
42. El promedio anual de gastos de una familia estadounidense en transporte diario es $6312 (*Money*, agosto de 2001). Suponga que dicha cantidad está distribuida normalmente.
    a. Si sabe que 5% de las familias estadounidenses gastan menos de $1000 en el transporte diario. ¿Cuál es la desviación estándar en esta cantidad de gasto?
    b. ¿Cuál es la probabilidad de que un hogar gaste entre $4000 y $6000?
    c. ¿En que intervalo se encuentran los gastos de las familias que constituyen 3% de las familias con los gastos más altos en transporte?
43. *Condé Nast Traveler* publica la lista de oro de los mejores hoteles en todo el mundo. Broadmoor Hotel en Colorado Springs tiene 700 habitaciones y estuvo en la lista de oro en 2004 (*Condé Nast Traveler*, enero de 2004). El grupo encargado del marketing de este hotel pronostica una demanda media de 670 habitaciones para el próximo fin de semana. Suponga que la demanda para el próximo fin de semana está distribuida normalmente y que la desviación estándar es 30.
    a. ¿Cuál la probabilidad de que se ocupen todas las habitaciones del hotel?
    b. ¿Cuál la probabilidad de que se ocupen 50 o más habitaciones del hotel?
    c. ¿Recomendaría al hotel hacer una promoción para aumentar la demanda? ¿Qué consideraciones serían importantes?
44. Ward Doering Auto Sales está pensando en ofrecer un contrato especial de servicio que cubra todos los costos de servicio de los automóviles vendidos. De acuerdo con la experiencia, el director de la empresa estima que los costos anuales de servicio están distribuidos casi normalmente con una media de $150 y una desviación estándar de $25.
    a. Si la empresa ofrece a los clientes el contrato de servicio por una cantidad anual de $200, ¿cuál es la probabilidad de que el costo de un servicio sea mayor a los $200 del precio del contrato?
    b. ¿Cuál es la ganancia esperada por la empresa en estos contratos de servicio?
45. ¿La falta de sueño es causa de accidentes de tráfico de consecuencias fatales? En un estudio se encontró que el número promedio por año de accidentes de tráfico con consecuencias fatales ocasionados por conductores somnolientos es 1550 (*BusinessWeek*, 26 de enero de 2004). Suponga que el número promedio anual de accidentes de tráfico de consecuencias fatales está distribuido normalmente con una desviación estándar de 300.
    a. ¿Cuál es la probabilidad de que haya menos de 1000 accidentes fatales en un año?
    b. ¿De que el número anual de accidentes fatales esté entre 1000 y 2000?
    c. Para que un año se encuentre en el 5% superior en número de accidentes fatales, cuántos de éstos tendrán que ocurrir?

46. Suponga que las puntuaciones obtenidas en el examen de admisión a una universidad están distribuidas en forma normal con una media de 450 y una desviación estándar de 100.
    a. ¿Qué porcentaje de las personas que hacen el examen tendrá una puntuación entre 400 y 500?
    b. Si la puntuación que obtiene un estudiante es 630. ¿Qué porcentaje de los estudiantes que hacen el examen tendrá una puntuación mayor? ¿Qué porcentaje tendrá una puntuación menor?
    c. Si la universidad no admite estudiantes que obtengan una puntuación menor a 480, ¿qué porcentaje de los estudiantes que hacen el examen podrá ser aceptado?
47. De acuerdo con *Adversiting Age*, el salario base promedio de las mujeres que trabajan como publicistas es superior al salario base promedio de los hombres. El salario base promedio de las mujeres es \$67 000 y el salario base promedio de los hombres es \$65 500 (*Working Woman*, julio/agosto de 2000). Suponga que los salarios están distribuidos normalmente con una desviación estándar de \$7000 tanto para hombres como para mujeres.
    a. ¿Cuál es la probabilidad de que una mujer tenga un salario mayor que \$75 000?
    b. ¿De que un hombre tenga un salario mayor que \$75 000?
    c. ¿De que una mujer tenga un salario mayor que \$50 000?
    d. ¿Cuánto tendrá que ganar una mujer para tener un salario mayor que 99% de los hombres?
48. Una máquina llena recipientes con un determinado producto. De acuerdo con datos anteriores se sabe que la desviación estándar en los pesos rellenados es 0.6 onzas. Si sólo 2% de los recipientes llenados tienen menos de 18 onzas, ¿cuál es el peso medio de llenado de la máquina? Es decir, a cuánto es igual μ? Suponga que los pesos llenados tienen una distribución normal.
49. Considere un examen de opción múltiple con 50 preguntas. Para cada pregunta hay cuatro respuestas posibles. Suponga que un estudiante que ha hecho las tareas y asistido a clase tiene 75% de probabilidad de contestar correctamente las preguntas.
    a. Para obtener A de calificación, un estudiante tiene que contestar correctamente 43 o más preguntas. ¿Qué porcentaje de los estudiantes que hicieron las tareas y asistieron a clase obtendrá A de calificación?
    b. Para obtener C de calificación, un estudiante tiene que contestar correctamente de 35 a 39 preguntas. ¿Qué porcentaje de los estudiantes que hicieron las tareas y asistieron a clases obtendrá C de calificación?
    c. Para aprobar el examen, un estudiante tiene que contestar correctamente 30 preguntas o más. ¿Qué porcentaje de los estudiantes que hicieron las tareas y asistieron a clases pasará el examen?
    d. Suponga que un estudiante no asistió a clases ni hizo las tareas. Además, dicho estudiante sólo tratará de adivinar las respuestas a las preguntas. ¿Cuál es la probabilidad de que el estudiante conteste correctamente 30 o más preguntas y pase el examen?
50. En Las Vegas un jugador de blackjack se entera de que la casa proporcionará una habitación gratis a quien juegue cuatro horas con un promedio de apuesta de \$50. La estrategia del jugador tiene una probabilidad de ganar en cualquier mano de 0.49 y el jugador sabe que hay 60 manos por hora. Suponga que el jugador juega durante cuatro horas con una apuesta de \$50 por mano.
    a. ¿Cuál es la ganancia esperada del jugador?
    b. ¿Cuál es la probabilidad de que el jugador pierda \$1000 o más?
    c. ¿Cuál es la probabilidad de que el jugador gane?
    d. Si el jugador empieza con \$1500. ¿Cuál es la probabilidad de que el jugador se vaya a la bancarrota?
51. El tiempo, en minutos, que un estudiante usa una terminal de computadora en el centro de cálculo de una universidad sigue una distribución de probabilidad exponencial con una media de 36 minutos. Suponga que un estudiante llega a una terminal precisamente en el momento en que otro estudiante quería usar la terminal.
    a. ¿Cuál es la probabilidad de que el segundo estudiante tenga que esperar 15 minutos o menos?
    b. ¿De que el segundo estudiante tenga que esperar entre 15 y 45 minutos?
    c. ¿Cuál la probabilidad de que el segundo estudiante tenga que esperar una hora o más?
52. El sitio Web de Bed and Breakfast Inns of North America (www.cimarron.net) recibe aproximadamente siete visitas por minuto (*Time*, septiembre de 2001). Suponga que el número de visitantes por minuto sigue una distribución de probabilidad de Poisson.

a. ¿Cuál es el tiempo medio entre las visitas a este sitio de la Web?
b. Muestre la función de densidad de probabilidad exponencial para los tiempos entre las visitas a este sitio.
c. ¿Cuál es la probabilidad de que nadie visite este sitio en un lapso de 1 minuto?
d. ¿Cuál es la probabilidad de que nadie visite este sitio en un lapso de 12 minutos?

53. En la ciudad de Nueva York el tiempo de recorrido promedio al trabajo es de 36.5 minutos.
a. Suponga que la distribución de probabilidad exponencial es aplicable y muestre la función de densidad de probabilidad para las duraciones de los recorridos al trabajo en Nueva York.
b. ¿Cuál es la probabilidad de que un neoyorquino típico necesite entre 20 y 40 minutos para transportarse a su trabajo?
c. ¿De que un neoyorquino típico necesite más de 40 minutos para transportarse a su trabajo?

54. El tiempo (en minutos) entre dos llamadas telefónicas en la oficina de solicitud de servicios de una aseguradora tiene la siguiente distribución de probabilidad exponencial.

$$f(x) = 0.50e^{-0.50x} \quad \text{para } x \geq 0$$

a. ¿Cuál es el tiempo medio entre las llamadas telefónicas?
b. ¿Cuál es la probabilidad de que pasen 30 segundos o menos entre llamadas telefónicas?
c. ¿De que pase 1 minuto o menos entre las llamadas telefónicas?
d. ¿Cuál es la probabilidad de que pasen 5 minutos o más sin que haya llamadas telefónicas?

Caso problema Specialty Toys

Specialty Toys, Inc. vende una gran variedad de nuevos e innovadores juguetes para niños. Los directivos saben que la época prenavideña es la mejor oportunidad para la introducción de un nuevo juguete, en esta época muchas personas buscan cosas novedosas para los regalos navideños. Cuando Specialty descubre un nuevo juguete con un buen potencial de mercado, elige alguna fecha en octubre para su lanzamiento.

Para contar con los juguetes en octubre Specialty hace los pedidos a sus proveedores en junio o julio de cada año. La demanda de juguetes para niños puede ser muy volátil. Si un nuevo juguete se pone de moda, la posibilidad de que se agote suele incrementar la demanda hasta niveles altos y se pueden obtener grandes ganancias. Sin embargo, un nuevo juguete también puede fracasar dejando a Specialty con un gran inventario que debe vender a precios reducidos. La interrogante más importante que enfrenta la empresa es decidir cuántas unidades comprar de un juguete nuevo para satisfacer la demanda. Si compra muy pocos, perderá ventas; si compra demasiados, las ganancias se reducirán por los precios bajos que tendrá que ofrecer en una liquidación.

En la próxima temporada Specialty desea introducir un juguete nuevo que se llama *El osito pronosticador del clima*. Esta variación de un osito de peluche que habla es fabricada por una empresa en Taiwan. Cuando un niño oprime la mano del osito, éste empieza a hablar. El osito tiene un barómetro que le ayuda, de acuerdo con el estado del tiempo, a elegir una de cinco frases que pronostican el estado del tiempo. Las frases van desde “¡Parece que es un bonito día! Que te diviertas” hasta “Parece que va a llover. No se te olvide llevar tu paraguas”. Pruebas realizadas con el producto indican que, aunque no es preciso, sus pronósticos del tiempo son sorprendentemente buenos. Varios de los directivos de Specialty opinan que los pronósticos del tiempo del osito son tan buenos como muchos de los pronósticos del tiempo que se dan en televisión.

Como ocurre con todos los productos, Specialty se enfrenta a la pregunta de cuántos ositos ordenar para la temporada siguiente. Las cantidades que sugieren los directivos son 15 000, 18 000, 24 000 o 28 000 unidades. El intervalo tan amplio en que se encuentran estas cantidades indica una considerable discrepancia en lo que se refiere al potencial de mercado. El equipo de directivos le solicita a usted un análisis de las probabilidades de terminar el inventario de acuerdo con diversas cantidades a comprar, así como una estimación del potencial de ganancias y su ayuda para hacer una recomendación de la cantidad que se debe comprar. Specialty espera vender *El osito pronosticador del clima* a $24 con base en un costo de $16 por unidad. Si hay inventario sobrante después de la temporada de las fiestas decembrinas, Specialty venderá las unidades res-

tantes a $5 cada una. Después de revisar las ventas anteriores de productos semejantes, los expertos de Specialty pronostican una demanda esperada de 20 000 unidades y 0.95 de probabilidad de que la demanda esté entre 10 000 y 30 000.

Informe administrativo

Elabore un informe sobre los puntos siguientes y recomiende la cantidad a comprar de *El osito pronosticador del clima*.

1. Use los pronósticos de ventas para describir una distribución de probabilidad normal que pueda servir para aproximar la distribución de la demanda. Dibuje la distribución y dé su media y su desviación estándar.
2. Calcule la probabilidad de terminar el inventario de acuerdo con las cantidades a comprar sugeridas por los miembros del equipo de directivos.
3. Calcule las ganancias proyectadas de acuerdo con las cantidades a comprar sugeridas por los miembros del equipo de directivos bajo tres escenarios: el peor de los casos, en el cual se venderán 10 000 unidades, en el caso más probable, en el cual se venderán 20 000 unidades y en el mejor de los casos en el cual se venderán 30 000 unidades.
4. Uno de los directivos de Specialty encuentra que el potencial de ganancia es tan bueno que la cantidad a comprar debe tener 70% de posibilidades de satisfacer la demanda y 30% de posibilidades de quedarse sin mercancía. De acuerdo con esto, ¿qué cantidad debe comprarse y cuál es la ganancia proyectada bajo cada uno de los tres escenarios?
5. Dé su propia recomendación sobre la cantidad que debe comprarse y muestre las proyecciones de ganancia correspondientes. Fundamente su recomendación.

Apéndice 6.1 Distribuciones de probabilidad continua con Minitab

Para demostrar el procedimiento de Minitab para el cálculo de probabilidades continuas se retomará el problema de la empresa Grear Tire, en el que la duración de los neumáticos en millas se describió mediante una distribución normal en la que $\mu = 36\,500$ y $\sigma = 5000$. Una de las preguntas que se plantearon fue: ¿cuál es la probabilidad de que los neumáticos duren más de 40 000 millas?

Para distribuciones de probabilidad continua, Minitab proporciona probabilidades acumuladas; es decir, Minitab da la probabilidad de que la variable aleatoria tome un valor menor o igual que una constante específica. En el caso de la pregunta sobre la duración de los neumáticos, Minitab se puede usar para determinar la probabilidad acumulada de que un neumático dure 40 000 millas o menos. (En este caso la constante específica es 40 000.) Una vez que se tiene la probabilidad acumulada que proporciona Minitab, es necesario restar esta probabilidad de 1 para determinar la probabilidad de que el neumático dure más de 40 000 millas.

Para que Minitab calcule una probabilidad, es necesario ingresar la constante específica en una de las columnas de la hoja de cálculo. En este caso se introduce la constante específica 40 000 en la columna C1 de la hoja de cálculo de Minitab. A continuación se presentan los pasos necesarios para que Minitab calcule la probabilidad acumulada de que la variable aleatoria normal tome valores menores o iguales que 40 000.

Paso 1. Seleccionar el menú **Calc**
Paso 2. Elegir **Probability Distributions**
Paso 3. Elegir **Normal**
Paso 4. Cuando aparezca el cuadro de diálogo Normal Distribution:
Seleccionar **Cumulative probability**
Ingresar 36 500 en el cuadro **Mean**
Ingresar 5 000 en el cuadro **Standard deviation**
Ingresar C1 en el cuadro **Input column** (la columna que contiene 40 000)
Clic en **OK**

Después de que el usuario hace clic en **OK**, Minitab da la probabilidad acumulada de que la variable aleatoria normal tome un valor menor o igual que 40 000. Minitab indica que esta probabilidad es 0.7580. Como lo que interesa es la probabilidad de que el neumático dure más de 40 000, la probabilidad buscada es $1 - 0.7580 = 0.2420$.

Otra pregunta en el problema de la empresa Grear Tire fue: ¿cuál es la duración en millas que la empresa debe establecer en la garantía de manera que en no más de 10% de los neumáticos se tenga que pagar la garantía? En este caso se da una probabilidad y se quiere hallar el valor correspondiente de la variable aleatoria. Minitab usa una rutina de cálculo inverso para hallar el valor de la variable aleatoria que corresponde a la probabilidad acumulada dada. Primero, se ingresa la probabilidad acumulada en la hoja de cálculo de Minitab (por ejemplo en C1). En este caso la probabilidad acumulada es 0.10. Después, los tres primeros pasos del procedimiento de Minitab son los dados antes. En el paso 4 se selecciona **Inverse cumulative probability** en lugar de **Cumulative probability** y se realiza la parte restante de este paso. Minitab da entonces 30 092 millas para la duración en la garantía.

Minitab también calcula las probabilidades de otras distribuciones de probabilidad continua, entre las que se encuentra la distribución de probabilidad exponencial. Para calcular probabilidades exponenciales se sigue el procedimiento antes dado para la distribución de probabilidad normal eligiendo la opción **Exponential** en el paso 3. El paso 4 es igual, salvo que no es necesario ingresar la desviación estándar. Minitab da los resultados de probabilidades acumuladas o probabilidades acumuladas inversas en la misma forma que se describió para la distribución de probabilidad normal.

Apéndice 6.2 Distribuciones de probabilidad continua con Excel

Excel permite calcular las probabilidades de varias distribuciones de probabilidad continuas. Entre las que se encuentran las distribuciones de probabilidad normal y exponencial. En este apéndice, se describe cómo usar Excel para calcular probabilidades en cualquier distribución normal. El procedimiento para la exponencial y para las otras distribuciones continuas es semejante al descrito aquí para la distribución normal.

Recuerde el problema de la empresa Grear Tire, la duración de los neumáticos en millas se describe mediante una distribución normal con media $\mu = 36\,500$ y $\sigma = 5000$. Suponga que se desea conocer la probabilidad de que un neumático dure más de 40 000 millas.

La función de Excel DISTR.NORM. suministra probabilidades acumuladas de una distribución normal. La forma general de la función es DISTR.NORM (x,media,desv_estándar,acum). En el cuarto argumento se especifica VERDADERO si se desea una probabilidad acumulada. De esta manera, para calcular la probabilidad acumulada de que la duración de un neumático sea menor o igual que 40 000 millas se ingresará la fórmula siguiente en cualquier celda de la hoja de cálculo Excel:

=DISTR.NORM(40000,36500,5000,VERDADERO)

En este momento, en la celda en que se ingresó la fórmula aparecerá 0.7580, indicando que la probabilidad de que la duración del neumático sea 40 000 millas es 0.7580. Por tanto, la probabilidad de que un neumático dure más de 40 000 millas es $1 - 0.7580 = 0.2420$.

La función de Excel DISTR.NORM.INV. usa un cálculo inverso para hallar el valor de x que corresponde a una probabilidad acumulada dada. Por ejemplo, si se desea hallar la duración que Grear debe ofrecer en su garantía de manera que no más de 10% de los neumáticos sean aptos para solicitar la garantía. Se ingresará en cualquier celda de la hoja de cálculo de Excel la fórmula siguiente:

=DISTR.NORM.INV(.1,36500,5000)

En este momento, en la celda en la que se ingresó la fórmula aparecerá 30092, indicando que la probabilidad de que un neumático dure 30 092 millas es 0.10.

La función de Excel para calcular probabilidades exponenciales es DISTR.EXP. Usar esta función es muy sencillo. Pero si se necesita ayuda para especificar los argumentos adecuados, se puede usar la herramienta Insertar Función de Excel (véase apéndice E).

CAPÍTULO 7



Muestreo y distribuciones muestrales

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

MEADWESTVACO CORPORATION*
STAMFORD, CONNECTICUT

MeadWestvaco Corporation, líder mundial en la producción de embalajes y papeles especiales, productos de consumo y de oficina y de sustancias químicas especiales, emplea a más de 30 000 personas. Opera a nivel mundial en 29 países y atiende a clientes localizados en 100 países. MeadWestvaco tiene una posición líder en la producción de papel, con una capacidad de 1.8 millones de toneladas anuales. Entre los productos de la empresa se encuentran papel para libros de texto, papel para revistas, sistemas de embalaje para bebidas y productos de oficina. Los consultores internos de MeadWestvaco usan el muestreo para obtener diversas informaciones que permiten a la empresa ganar productividad y seguir siendo competitiva.

Por ejemplo, MeadWestvaco posee bosques que le proporcionan los árboles, o la materia prima, para muchos de los productos de la empresa. Los directivos necesitan información confiable y precisa acerca de los bosques maderables para evaluar las posibilidades de satisfacción de las futuras necesidades de materia prima. ¿Cuál es el volumen actual de los bosques? ¿Cuál ha sido el crecimiento de los bosques? ¿Cuál es el crecimiento proyectado de los bosques? Las respuestas a estas preguntas permiten a los directivos de la empresa elaborar los planes para el futuro, tales como planes a largo plazo y calendarios para la poda de árboles.

¿Cómo recolecta MeadWestvaco la información que necesita acerca de los amplios bosques que requiere? Los datos que obtiene de puntos muestrales en los bosques son la base para contar con información acerca de la población de árboles propiedad de la empresa. Para localizar estos puntos muestrales, primero se dividen los bosques en tres secciones de acuerdo con la localización y tipo de árboles. Mediante mapas y números aleatorios los analistas de MeadWestvaco identifican puntos muestrales aleatorios de 1/5 a 1/7 acres en cada sección del bosque. Los ingenieros forestales de MeadWestvaco recogen los datos de estos puntos muestrales para obtener información acerca de la población forestal.



El muestreo aleatorio de sus bosques permite a MeadWestvaco satisfacer necesidades futuras de materia prima. © Walter Hodges/Corbis.

También participan en el proceso de campo de la recolección de datos. Con periodicidad, en equipos de dos personas, recolectan la información de cada árbol en todos los puntos muestrales. Los datos muestrales se ingresan en el sistema computacional de inventario forestal continuo (IFC) de la empresa. Los informes obtenidos del sistema IFC contienen información de distribuciones de frecuencia con estadísticos sobre los tipos de árboles, volumen de los bosques, tasas de crecimiento anteriores y crecimiento y volumen proyectados para el futuro. El muestreo y las correspondientes informaciones estadísticas de los datos muestrales proporcionan la información esencial para la adecuada administración de los bosques de MeadWestvaco.

En este capítulo se estudiará el muestreo aleatorio simple y el proceso de selección de muestras. Se verá también cómo se usan estadísticos como la media muestral y la proporción muestral para estimar la media de la población y la proporción de la población.

*Los autores agradecen al doctor Edgard P. Winkofsky por proporcionar la información para *La estadística en la práctica*.

En el capítulo 1 se definieron los términos población y muestra. Estas definiciones se retoman aquí.

1. Una *población* es el conjunto de todos los elementos que interesan en un estudio.
2. Una *muestra* es un subconjunto de la población.

A las características numéricas de una población, como la media y la desviación estándar, se les llama **parámetros**. El principal propósito de la inferencia estadística es hacer estimaciones y pruebas de hipótesis acerca de los parámetros poblacionales usando la información que propor-

ciona una muestra. Para empezar, se presentan dos situaciones en las que a partir de muestras se obtienen estimaciones de parámetros poblacionales.

1. Un fabricante de neumáticos elabora un nuevo modelo que tendrá mayor duración que los actuales neumáticos de la empresa. Para estimar la duración media, en millas, el fabricante selecciona una muestra de 120 neumáticos nuevos para probarlos. De los resultados de esta prueba se obtiene una duración media de 36 500 millas. Por tanto, una estimación de la duración media, en millas, de la población de nuevos neumáticos es 36 500 millas.
2. Los miembros de un partido político deseaban apoyar a un determinado candidato para senador, y los dirigentes del partido deseaban tener una estimación de la proporción de votantes registrados que podían estar a favor del candidato. El tiempo y el costo de preguntar a cada uno de los individuos de la población de votantes registrados eran prohibitivos. Por tanto, se seleccionó una muestra de 400 votantes registrados; 160 de los 400 votantes indicaron estar a favor del candidato. Una estimación de la proporción de la población de votantes registrados a favor del candidato es 160/400 = 0.40.

Estos dos ejemplos ilustran algunas de las razones por las que se usan muestras. Observe que en el ejemplo de los neumáticos, obtener datos sobre su tiempo de duración implica usarlos hasta que se acaben. Es claro que no es posible probar toda la población de neumáticos; una muestra es la única manera factible de obtener los datos de duración deseados. En el ejemplo del candidato, preguntar a cada uno de los votantes registrados es, en teoría, posible, pero el tiempo y el costo para hacerlo son prohibitivos; de manera que se prefiere una muestra de los votantes registrados.

Una media muestral suministra una estimación de la media poblacional y una proporción muestral suministra una estimación de la proporción poblacional. Con dichas estimaciones puede esperarse un cierto error de estimación. Este capítulo enseña las bases para estimar cuán grande puede ser ese error.

Es importante darse cuenta de que los resultados muestrales sólo proporcionan una *estimación* de los valores de las características de la población. No se espera que la media muestral de 36 500 millas sea exactamente igual al millaje medio de todos los neumáticos de la población, tampoco que 0.40, o 40% de la población de los votantes registrados esté a favor del candidato. La razón es simple, la muestra sólo contiene una parte de la población. Con métodos de muestreo adecuados, los resultados muestrales proporcionarán estimaciones "buenas" de los parámetros poblacionales. Pero ¿cuán buenos puede esperarse que sean los resultados muestrales? Por fortuna, existen procedimientos estadísticos para responder esta pregunta.

En este capítulo se enseña cómo emplear el muestreo aleatorio simple para seleccionar una muestra de una población. Después, cómo usar una muestra aleatoria simple para calcular estimaciones de una media poblacional, de una desviación estándar poblacional y de una proporción poblacional. Además, también se presenta el importante concepto de distribución muestral. Como verá, el conocimiento de la distribución muestral adecuada permite decir qué tan cerca se encuentran las estimaciones muestrales de los correspondientes parámetros poblacionales. En la última sección se estudian alternativas al muestreo aleatorio simple, usadas con frecuencia en la práctica.

7.1 El problema de muestreo de Electronics Associates

Al director de personal de Electronics Associates, Inc. (EAI), se le ha encargado la tarea de elaborar un perfil de los 2500 administradores de la empresa. Las características a determinar son el sueldo medio anual de los administradores y la proporción de administradores que ha terminado el programa de capacitación de la empresa.



EAI

Con los 2500 administradores de la empresa como la población para este estudio, es posible hallar el sueldo anual y la situación respecto al programa de capacitación de cada persona al consultar los archivos del personal. El archivo con los datos que contiene esta información para cada uno de los 2500 administradores que forman la población se encuentra en el disco compacto que se distribuye con el libro.

Con los datos de EAI y las fórmulas presentadas en el capítulo 3, se calcula la media poblacional y la desviación estándar poblacional de los salarios anuales.

$$\text{Media poblacional:} \quad \mu = \$51\,800$$
$$\text{Desviación estándar poblacional:} \quad \sigma = \$4000$$

Los datos sobre la situación del programa de capacitación muestran que 1500 de los 2500 administradores han terminado el programa de capacitación. Si p denota la proporción de la población que ha terminado el programa de capacitación, se tiene que $p = 1500/2500 = 0.60$. La media poblacional de los sueldos anuales ($\mu = \$51\ 800$), la desviación estándar poblacional de los sueldos anuales ($\sigma = \$4000$) y la proporción poblacional de quienes han terminado el programa de capacitación ($p = 0.60$) son parámetros de la población de administradores de EAI.

Algunos de los costos de recopilar información de una muestra son sustancialmente menores que hacerlo de una población; especialmente cuando se deben realizar entrevistas personales para recopilar la información.

Ahora suponga que la información necesaria sobre todos los administradores de EAI no esté disponible en la base de datos de la empresa. La pregunta que se considera ahora es: ¿cómo puede obtener el director de personal de la empresa, estimaciones de los parámetros poblacionales usando una muestra de los administradores, en lugar de usar a los 2500 administradores de la población. Asuma que se va a emplear una muestra de 30 administradores. Es obvio que el tiempo y el costo de la elaboración de un perfil será mucho menor usando 30 administradores que la población entera. Si el director de personal tuviera la certeza de que una muestra de 30 administradores proporciona la información adecuada acerca de la población de los 2500 administradores, preferiría trabajar con una muestra que hacerlo con toda la población. Para explorar la posibilidad de usar una muestra para el estudio de EAI, primero se considerará cómo determinar una muestra de 30 administradores.



7.2 Muestreo aleatorio simple

Para seleccionar una muestra de una población hay diversos métodos; uno de los más comunes es el **muestreo aleatorio simple**. La definición de muestreo aleatorio simple y del proceso de seleccionar una muestra aleatoria simple dependen de si la población es *finita* o *infinita.* Como el problema de muestreo de EAI tiene una población finita de 2500 administradores, primero se considera el muestreo de una población finita.

Muestreo de una población finita

Una muestra aleatoria simple de tamaño n de una población finita de tamaño N se define como sigue.

MUESTREO ALEATORIO SIMPLE (POBLACIÓN FINITA)

Una muestra aleatoria simple de tamaño n de una población finita de tamaño N es una muestra seleccionada de manera que cada posible muestra de tamaño n tenga la misma probabilidad de ser seleccionada.

Un procedimiento para seleccionar una muestra aleatoria simple de una población finita es elegir los elementos para la muestra de uno en uno, de manera que, en cada paso, cada uno de los elementos que quedan en la población tenga la misma probabilidad de ser seleccionado. Al seleccionar n elementos de esta manera, será satisfecha la definición de muestra aleatoria simple seleccionada de una población finita.

Los números aleatorios generados por computadora también sirven para realizar el proceso de selección de una muestra aleatoria. Excel proporciona una función para generar números aleatorios en sus hojas de cálculo.

Los números aleatorios en la tabla aparecen en grupos de cinco para facilitar su lectura.

Para seleccionar una muestra aleatoria simple de la población finita de administradores de EAI, primero se le asigna a cada administrador un número. Por ejemplo, se les asignan los números del 1 al 2500 en el orden en que aparecen sus nombres en el archivo de personal de EAI. A continuación se consulta la tabla de dígitos aleatorios que se muestran en la tabla 7.1. Al consultar el primer renglón de la tabla se da cuenta que cada dígito, 6, 3, 2, … es un dígito aleatorio con la misma oportunidad de aparecer que cualquier otro. Como el número mayor en la lista de la población de administradores de EAI, 2500, tiene cuatro dígitos, se seleccionarán números aleatorios de la tabla en conjuntos o grupos de cuatro dígitos. Aun cuando para la selección de números aleatorios se puede empezar en cualquier lugar de la tabla y avanzar sistemáticamente en una de las cuatro direcciones, aquí se usará el primer renglón de la tabla 7.1 y se avanzará de izquierda a derecha. Los primeros 7 números aleatorios de cuatro dígitos son

6327 1599 8671 7445 1102 1514 1807

Los números aleatorios en la tabla aparecen en grupos de cinco para facilitar su lectura.

TABLA 7.1 NÚMEROS ALEATORIOS

63271	59986	71744	51102	15141	80714	58683	93108	13554	79945
88547	09896	95436	79115	08303	01041	20030	63754	08459	28364
55957	57243	83865	09911	19761	66535	40102	26646	60147	15702
46276	87453	44790	67122	45573	84358	21625	16999	13385	22782
55363	07449	34835	15290	76616	67191	12777	21861	68689	03263
69393	92785	49902	58447	42048	30378	87618	26933	40640	16281
13186	29431	88190	04588	38733	81290	89541	70290	40113	08243
17726	28652	56836	78351	47327	18518	92222	55201	27340	10493
36520	64465	05550	30157	82242	29520	69753	72602	23756	54935
81628	36100	39254	56835	37636	02421	98063	89641	64953	99337
84649	48968	75215	75498	49539	74240	03466	49292	36401	45525
63291	11618	12613	75055	43915	26488	41116	64531	56827	30825
70502	53225	03655	05915	37140	57051	48393	91322	25653	06543
06426	24771	59935	49801	11082	66762	94477	02494	88215	27191
20711	55609	29430	70165	45406	78484	31639	52009	18873	96927
41990	70538	77191	25860	55204	73417	83920	69468	74972	38712
72452	36618	76298	26678	89334	33938	95567	29380	75906	91807
37042	40318	57099	10528	09925	89773	41335	96244	29002	46453
53766	52875	15987	46962	67342	77592	57651	95508	80033	69828
90585	58955	53122	16025	84299	53310	67380	84249	25348	04332
32001	96293	37203	64516	51530	37069	40261	61374	05815	06714
62606	64324	46354	72157	67248	20135	49804	09226	64419	29457
10078	28073	85389	50324	14500	15562	64165	06125	71353	77669
91561	46145	24177	15294	10061	98124	75732	00815	83452	97355
13091	98112	53959	79607	52244	63303	10413	63839	74762	50289

Como los números de la tabla son aleatorios, estos números de cuatro dígitos son todos igualmente posibles. Ahora se pueden usar estos números aleatorios de cuatro dígitos para darle a cada uno de los administradores que constituyen la población la misma oportunidad de ser incluido en la muestra aleatoria. El primer número, 6327, es mayor que 2500. No corresponde a ninguno de los administradores numerados que forman la población y por tanto se descarta. El segundo número, 1599, está entre 1 y 2500. Por tanto, el primer administrador seleccionado para la muestra aleatoria es el administrador que tiene el número 1599 en la lista de los administradores de EAI. Siguiendo este proceso, se ignoran los números 8671 y 7445 antes de identificar a los administradores con los números 1102, 1514 y 1807 e incluirlos en la muestra aleatoria. Este proceso sigue hasta que se tiene la muestra aleatoria de 30 administrativos de EAI.

Al realizar este proceso para la selección de una muestra aleatoria simple, es posible que un número que ya haya sido usado se encuentre de nuevo en la tabla antes de completar la muestra de los 30 administradores de EAI. Como no se quiere seleccionar a un administrador más de una vez, cualquier número aleatorio que ya ha sido usado se ignora, porque el administrador correspondiente ya se ha incluido en la muestra. A este tipo de selección se le conoce como **muestreo sin reemplazo**. Cuando se selecciona una muestra en la que se acepten números aleatorios ya usados y los administradores correspondientes sean incluidos dos o más veces, se está **muestreando con reemplazo**. Muestrear con reemplazo es una forma válida de identificar una muestra aleatoria simple. Sin embargo, el muestreo sin reemplazo es el procedimiento de muestreo más usado. Cuando se habla de muestreo aleatorio simple, se asumirá que el muestreo es sin reemplazo.

En la práctica, la población en estudio se considera infinita si se tiene un proceso continuo en el que sea imposible contar o enumerar cada uno de los elementos de la población.

Muestreo de una población infinita

En algunas situaciones la población o bien es infinita o tan grande que, para fines prácticos, se considera infinita. Por ejemplo, suponga que un restaurante de comida rápida desea obtener el

perfil de su clientela seleccionando una muestra aleatoria de los mismos y pidiéndole a cada cliente que llene un breve cuestionario. En tales situaciones, el proceso continuo de clientes que visitan el restaurante puede verse como que los clientes provienen de una población infinita. La definición de muestra aleatoria simple tomada de una población infinita es la siguiente

MUESTRA ALEATORIA SIMPLE (POBLACIÓN INFINITA)

Una muestra aleatoria simple de una población infinita es una muestra seleccionada de manera que se satisfagan las condiciones siguientes.

1. Cada uno de los elementos seleccionados proviene de la población.
2. Cada elemento se selecciona independientemente.

En poblaciones infinitas un procedimiento para la selección de una muestra debe ser concebido especialmente para cada situación, de manera que permita seleccionar los elementos de manera independiente y evitar así un sesgo en la selección que dé mayores probabilidades de selección a ciertos tipos de elementos.

En poblaciones infinitas un procedimiento para la selección de una muestra debe ser concebido especialmente para cada situación, de manera que permita seleccionar los elementos de manera independiente y evitar así un sesgo en la selección que dé mayores probabilidades de selección a ciertos tipos de elementos. En el ejemplo de la selección de una muestra aleatoria simple entre los clientes de un restaurante de comida rápida, el primer requerimiento es satisfecho por cualquier cliente que entra en el restaurante. El segundo requerimiento es satisfecho seleccionando a los clientes de manera independiente. El objetivo del segundo requerimiento es evitar sesgos de selección. Habría un sesgo de selección, por ejemplo, si cinco clientes consecutivos que se seleccionaran fueran amigos. Es de esperar que estos clientes tengan perfiles semejantes. Dichos sesgos se evitan haciendo que la selección de un cliente no influya en la selección de cualquier otro cliente. En otras palabras, los clientes deben ser seleccionados de manera independiente.

McDonald´s, el restaurante líder en comida rápida, realizó un muestreo aleatorio simple precisamente en una situación así. El procedimiento de muestreo se basó en el hecho de que algunos clientes presentaban cupones de descuento. Cada vez que un cliente presentaba un cupón descuento, al siguiente cliente que se atendía se le pedía que llenara un cuestionario sobre el perfil del cliente. Como los clientes que llegaban al restaurante presentaban cupones de descuento aleatoria e independientemente, este plan de muestreo garantizaba que los clientes fueran seleccionados de manera independiente. Por tanto, los dos requerimientos para un muestreo aleatorio simple de una población infinita fueron satisfechos.

Las poblaciones infinitas suelen asociarse con un proceso que opera continuamente a lo largo del tiempo. Por ejemplo, partes fabricadas en una línea de producción, transacciones en un banco, llamadas que llegan a un centro de asesoría técnica y clientes que entran en las tiendas son considerados como provenientes de una población infinita. En tales casos un procedimiento de muestreo creativo garantiza que no haya sesgos de selección y que los elementos de la muestra sean seleccionados en forma independiente.

NOTAS Y COMENTARIOS

1. El número de muestras aleatorias simples distintas de tamaño n que pueden seleccionarse de una población finita de tamaño N es

$$\frac{N!}{n!(N-n)!}$$

En esta fórmula $N!$ y $n!$ son factoriales, vistos en el capítulo 4. Al usar esta expresión con los datos del problema de EAI, en el que $N = 2500$ y $n = 30$, se ve que se pueden tomar 2.75×10^{69} muestras aleatorias simples distintas de 30 administradores de EAI.

2. Para tomar una muestra aleatoria pueden emplearse paquetes de software. En los apéndices del capítulo se muestra cómo usar Minitab y Excel para seleccionar una muestra aleatoria simple de una población finita

Ejercicios

Método



1. Dada una población finita que tiene cinco elementos A, B, C, D y E seleccione 10 muestras aleatorias simples de tamaño 2.
   a. Enumere las 10 muestras empezando con AB, AC y así en lo sucesivo.
   b. Usando el muestreo aleatorio simple, ¿cuál es la probabilidad que tiene cada muestra de tamaño 2 de ser seleccionada?
   c. Si el número aleatorio 1 corresponde a A, el número 2 corresponde a B y así en lo sucesivo. Enliste la muestra aleatoria de tamaño 2 que será seleccionada al usar los números aleatorios 8 0 5 7 5 3 2.
2. Suponga que una población finita tiene 350 elementos. A partir de los últimos tres dígitos de cada uno de los siguientes números aleatorios de cinco dígitos (por ejemplo: 601, 022, 448,...), determine los primeros cuatro elementos que se seleccionarán para una muestra aleatoria simple.

   98601 73022 83448 02147 34229 27553 84147 93289 14209

Aplicaciones



3. *Fortune* publicó datos sobre ventas, valor del activo, valor de mercado y ganancias por acción de las 500 corporaciones industriales más grandes de Estados Unidos (*Fortune* 500, 2003). Suponga que usted desea seleccionar una muestra aleatoria simple de 10 corporaciones de la lista *Fortune* 500. Use los tres últimos dígitos de la columna 9 de la tabla 7.1, empezando con 554. Leyendo hacia abajo por esa columna, identifique los números de las 10 corporaciones que se tomarán para la muestra.
4. A continuación se presentan las 10 acciones más activas en la Bolsa de Nueva York del 6 de marzo del 2006 (*The Wall Street Journal,* 7 de marzo, 2006).

AT&T	Lucent	Nortel	Qwest	Bell South
Pfizer	Texas Instruments	Gen. Elect.	iShrMSJpn	LSI Logic

   Las autoridades decidieron investigar las prácticas de negociación usando una muestra de tres de estas acciones.
   a. Empezando en el primer dígito aleatorio de la columna seis de la tabla 7.1, lea los números descendiendo por esa columna para seleccionar una muestra aleatoria simple de tres acciones para las autoridades.
   b. Con la información dada en la primera nota y comentario, determine cuántas muestras aleatorias simples diferentes de tamaño 3 pueden seleccionarse de una lista de 10 acciones.
5. Una organización de estudiantes desean estimar la proporción de estudiantes que están a favor de una disposición de la escuela. Se cuenta con una lista con los nombres y direcciones de los 645 estudiantes inscritos el presente trimestre. Tomando números aleatorios de tres dígitos del renglón 10 de la tabla 7.1 y avanzando por ese renglón de izquierda a derecha, determine los 10 primeros estudiantes que serán seleccionados usando un muestreo aleatorio simple. Los números aleatorios de tres dígitos empiezan con 816, 283 y 610.
6. El *County and City Data Book* del Census Bureau cuenta con información de los 3139 condados de Estados Unidos. Suponga que para un estudio nacional se recogerán datos de 30 condados seleccionados aleatoriamente. De la última columna de la tabla 7.1 extraiga números aleatorios de cuatro dígitos para determinar los primeros cinco condados seleccionados para la muestra. Ignore los primeros dígitos y empiece con los números aleatorios de cuatro dígitos 9945, 8364, 5702 y así sucesivamente.
7. Suponga que se va a tomar una muestra aleatoria simple de 12 de los 372 médicos de una determinada ciudad. Una organización médica le proporciona los nombres de los médicos. De la tabla

7.1 use la columna ocho de números aleatorios de cinco dígitos para determinar cuáles serán los 12 médicos para la muestra. Ignore los primeros dos dígitos de cada grupo de cinco dígitos de números aleatorios. Este proceso empieza con el número aleatorio 108 y continúa descendiendo por la columna de números aleatorios.

8. La lista siguiente proporciona los 25 mejores equipos de futbol de la NCAA en la temporada del 2002 (*NCAA News*, 4 de enero de 2003). De la tabla 7.1 use la novena columna que empieza con 13 554, para seleccionar una muestra aleatoria simple de seis equipos de futbol. Empiece con el equipo 13 y use los primeros dos dígitos de cada renglón de la novena columna para el proceso de selección. ¿Cuáles son los seis equipos de futbol seleccionados para la muestra aleatoria simple?

1. Ohio State	14. Virginia Tech
2. Miami	15. Penn State
3. Georgia	16. Auburn
4. Southern California	17. Notre Dame
5. Oklahoma	18. Pittsburgh
6. Kansas State	19. Marshall
7. Texas	20. West Virginia
8. Iowa	21. Colorado
9. Michigan	22. TCU
10. Washington State	23. Florida State
11. North Carolina State	24. Florida
12. Boise State	25. Virginia
13. Maryland	

9. *The Wall Street Journal* proporciona el valor de activo neto, el rendimiento porcentual en lo que va del año y el rendimiento porcentual en tres años de 555 fondos mutualistas (*The Wall Street Journal*, 25 de abril de 2003). Suponga que se va a usar una muestra aleatoria simple de 12 de estos 555 fondos mutualistas para un estudio acerca de su tamaño y desempeño. Use la cuarta columna de números aleatorios en la tabla 7.1 empezando con el número 51102, para seleccionar la muestra aleatoria simple de 12 fondos mutualistas. Empiece con el fondo 102 y use los *últimos* tres dígitos de cada renglón de la cuarta columna para el proceso de selección. ¿Cuáles son los números de los 12 fondos mutualistas en esta muestra aleatoria simple?

10. Indique si las poblaciones siguientes se consideran finitas o infinitas.
    a. Todos los votantes registrados en el estado de California.
    b. Todos los equipos de televisión que pueden ser producidos en una determinada fábrica.
    c. Todas las órdenes que pueden ser procesadas por Allentown, Pensilvania, planta de TV-M Company.
    d. Todas las llamadas de emergencia que pueden ser recibidas en una estación de policía.
    e. Todas las piezas producidas por Fibercon, Inc., en el segundo turno el 17 de mayo.

7.3 Estimación puntual

Una vez descrito cómo seleccionar una muestra aleatoria simple, se vuelve al problema de EAI. En la tabla 7.2 se presenta una muestra aleatoria simple de 30 administradores con sus respectivos datos de sueldo anual y de participación en el programa de capacitación. La notación x_1, x_2, etc., se usa para denotar el sueldo anual del primer administrador de la muestra, del segundo, y así sucesivamente. La participación en el programa de capacitación se indica por un Sí en la columna programa de entrenamiento.

Para estimar el valor de un parámetro poblacional, la característica correspondiente se calcula con los datos de la muestra, a lo que se le conoce como **estadístico muestral**. Por ejemplo, para estimar la media poblacional μ y la desviación estándar poblacional σ de los salarios anuales de los administradores de EAI, se emplean los datos de la tabla 7.2 y se calculan los es-

TABLA 7.2 SALARIOS ANUALES Y SITUACIÓN RESPECTO AL PROGRAMA DE CAPACITACIÓN DE LOS ADMINISTRADORES PERTENECIENTES A UNA MUESTRA ALEATORIA SIMPLE DE 30 ADMINISTRADORES DE EAI

Salario anual	Programa de capacitación	Salario anual ($)	Programa de capacitación
x_1 = 49 094.30	Sí	x_{16} = 51 766.00	Sí
x_2 = 53 263.90	Sí	x_{17} = 52 541.30	No
x_3 = 49 643.50	Sí	x_{18} = 44 980.00	Sí
x_4 = 49 894.90	Sí	x_{19} = 51 932.60	Sí
x_5 = 47 621.60	No	x_{20} = 52 973.00	Sí
x_6 = 55 924.00	Sí	x_{21} = 45 120.90	Sí
x_7 = 49 092.30	Sí	x_{22} = 51 753.00	Sí
x_8 = 51 404.40	Sí	x_{23} = 54 391.80	No
x_9 = 50 957.70	Sí	x_{24} = 50 164.20	No
x_{10} = 55 109.70	Sí	x_{25} = 52 973.60	No
x_{11} = 45 922.60	Sí	x_{26} = 50 241.30	No
x_{12} = 57 268.40	No	x_{27} = 52 793.90	No
x_{13} = 55 688.80	Sí	x_{28} = 50 979.40	Sí
x_{14} = 51 564.70	No	x_{29} = 55 860.90	Sí
x_{15} = 56 188.20	No	x_{30} = 57 309.10	No

tadísticos muestrales correspondientes; media muestral $\bar{x}$ y desviación estándar muestral s. Con las fórmulas para la media muestral y la desviación estándar muestral presentadas en el capítulo 3 se obtiene que la media muestral es

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1\,554\,420}{30} = \$51\,814$$

y la desviación estándar muestral es

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{325\,009\,260}{29}} = \$3348$$

Para estimar p, la proporción de administradores que han terminado el programa de capacitación, se usa la proporción muestral correspondiente $\bar{p}$. Sea x el número de administradores de la muestra que han terminado el programa de capacitación. De acuerdo con la tabla 7.2, $x = 19$. Por tanto, como el tamaño de la muestra es $n = 30$, la proporción muestral es

$$\bar{p} = \frac{x}{n} = \frac{19}{30} = 0.63$$

Al hacer los cálculos anteriores, se lleva a cabo el proceso estadístico conocido como *estimación puntual.* A la media muestral $\bar{x}$ se le conoce como el **estimador puntual** de la media poblacional μ, a la desviación estándar muestral s como el estimador puntual de la desviación estándar poblacional σ y a la proporción muestral $\bar{p}$ como el estimador puntual de la proporción poblacional p. Al valor numérico obtenido de $\bar{x}$, s, o $\bar{p}$ se les conoce como **estimaciones puntuales**. Así, en la muestra aleatoria simple de 30 administradores de EAI que se presenta en la tabla 7.2, \$51 814 es la estimación puntual de μ, \$3 348 es la estimación puntual de σ y 0.63 es la estimación puntual de p. En la tabla 7.3 se resumen los resultados muestrales y se comparan las estimaciones puntuales con los valores de los parámetros poblacionales.

TABLA 7.3 INFORMACIÓN DE LAS ESTIMACIONES PUNTUALES OBTENIDAS DE UNA MUESTRA ALEATORIA SIMPLE DE 30 ADMINISTRADORES DE EAI

Parámetro poblacional	Valor del parámetro	Estimador puntual	Estimación puntual
μ = Media poblacional de los salarios anuales	\$51 800	$\bar{x}$ = Media muestral de los salarios anuales	\$51 814
σ = Desviación estándar poblacional de los salarios anuales	\$4 000	s = Desviación estándar muestral de los salarios anuales	\$3 348
p = Proporción poblacional que ha terminado el programa de capacitación	0.60	$\bar{p}$ = Proporción muestral que ha terminado el programa de capacitación	0.63

Como se observa en la tabla 7.3, las estimaciones puntuales difieren un poco de los correspondientes parámetros poblacionales. Estas diferencias son de esperarse ya que para elaborar las estimaciones muestrales se usa una muestra, y no un censo de toda la población. En el capítulo siguiente se verá cómo elaborar un intervalo de estimación para tener información acerca de qué tan cerca está la estimación muestral del parámetro poblacional.

Ejercicios

Métodos



11. Los datos siguientes provienen de una muestra aleatoria simple.

5 8 10 7 10 14

a. ¿Cuál es la estimación puntual de la media poblacional?
b. ¿Cuál es la estimación puntual de la desviación estándar poblacional?

12. Como respuestas a una pregunta de una encuesta a 150 individuos de una muestra se obtuvieron 75 Sí, 55 No y 20 individuos no dieron su opinión.
a. ¿Cuál es la estimación puntual de la proporción de la población que responde Sí?
b. ¿Cuál es la estimación puntual de la proporción de la población que responde No?

Aplicaciones



13. La siguiente información son datos obtenidos en una muestra aleatoria de las ventas de 5 meses:



Mes:	1	2	3	4	5
Unidades vendidas:	94	100	85	94	92

a. Calcule una estimación puntual de la media poblacional del número medio de unidades vendidas por mes.
b. Calcule una estimación puntual de la desviación estándar del número de unidades vendidas por mes.

archivo en CD
MutualFund

14. *BusinessWeek* publicó información sobre 283 fondos mutualistas (*BusinessWeek* 26 de enero de 2004). En el conjunto de datos MutualFunds se encuentra una muestra de 40 de estos fondos. Use este conjunto de datos para hacer lo que se pide en los incisos siguientes.
a. Calcule una estimación puntual de la proporción de fondos de inversión de *BusinessWeek* que son fondos de cargo.
b. Calcule una estimación puntual de la proporción de fondos clasificados como de alto riesgo.
c. Calcule una estimación puntual de la proporción de fondos con una puntuación abajo del promedio para el riesgo.

15. Muchos de los medicamentos empleados en el tratamiento del cáncer son costosos. *BusinessWeek* informó de los costos de los tratamientos con Herceptin, un medicamento para tratar el cáncer de

mama (*BusinessWeek*, 30 de enero de 2006). Los siguientes son los costos de tratamientos con Herceptin en una muestra aleatoria de 10 pacientes.

4376	5578	2717	4920	4495
4798	6446	4119	4237	3814

a. Calcule una estimación puntual del costo medio de un tratamiento con Herceptin.
b. Calcule una estimación puntual de la desviación estándar en los costos de los tratamientos con Herceptin.

16. En una muestra de 50 empresas de *Fortune* 500, 5 se encontraban en Nueva York, 6 en California, 2 en Minesota y 1 en Wisconsin.
    a. Dé una estimación de la proporción de empresas de *Fortune* 500 que se encuentran en Nueva York.
    b. Dé una estimación del número de empresas de *Fortune* 500 que se encuentran en Minesota.
    c. Dé una estimación de la proporción de empresas de *Fortune* 500 que no se encuentran en ninguno de estos estados.
17. La American Association of Individuals Investors (AAII) hace sondeos semanales entre sus suscriptores para determinar cuántos se muestran optimistas, pesimistas o indiferentes respecto al mercado de acciones a corto plazo. Sus hallazgos en la semana que terminó el 2 de marzo de 2006 son consistentes con los resultados muestrales siguientes (www.aaii.com).

    Optimistas 409 Indiferentes 299 Pesimistas 291

    Dé una estimación puntual de los parámetros poblacionales siguientes.
    a. Proporción de suscriptores de AAII optimistas respecto al mercado de acciones.
    b. Proporción de suscriptores de AAII indiferentes respecto al mercado de acciones.
    c. Proporción de suscriptores de AAII pesimistas respecto al mercado de acciones.

7.4 Introducción a las distribuciones muestrales

En la sección anterior se dijo que la media muestral $\bar{x}$ es el estimador puntual de la media poblacional μ y que la proporción muestral $\bar{p}$ es el estimador puntual de la proporción poblacional p. En la muestra aleatoria simple de los 30 administradores de EAI que se presenta en la tabla 7.2, la estimación puntual de μ es $\bar{x} = \$51\ 814$ y la estimación puntual de p es $\bar{p} = 0.63$. Suponga que se selecciona otra muestra aleatoria simple de 30 administradores de EAI y se obtienen las estimaciones puntuales siguientes:

$$\text{Media muestral: } \bar{x} = \$52\ 670$$
$$\text{Proporción muestral: } \bar{p} = 0.70$$

Observe que se obtuvieron valores diferentes de $\bar{x}$ y de $\bar{p}$. En efecto, otra muestra aleatoria simple de 30 administradores de EAI no se puede esperar que dé las mismas estimaciones puntuales que la primera muestra.

Ahora suponga que el proceso de seleccionar una muestra aleatoria simple de 30 administradores se repite una y otra y otra vez y que cada vez se calculan los valores de $\bar{x}$ y de $\bar{p}$. En la tabla 7.4 se muestra una parte de los resultados obtenidos en 500 muestras aleatorias simples y en la tabla 7.5 las distribuciones de frecuencias y distribuciones de frecuencias relativas de los valores de las 500 $\bar{x}$. En la figura 7.1 se muestra el histograma de las frecuencias de los valores de $\bar{x}$.

Poder entender el material de los capítulos siguientes depende de entender y usar las distribuciones muestrales que se presentan en este capítulo.

En el capítulo 5 se definió una variable aleatoria como una descripción numérica del resultado de un experimento. Si el proceso de seleccionar una muestra aleatoria simple se considera como un experimento, la media muestral $\bar{x}$ es el valor numérico del resultado de ese experimento. Por tanto, la media muestral es una variable aleatoria. Entonces, como ocurre con otras variables aleatorias, $\bar{x}$ tiene una media o valor esperado, una desviación estándar y una distribución

TABLA 7.4 VALORES DE $\bar{x}$ Y DE $\bar{p}$ OBTENIDOS EN 500 MUESTRAS ALEATORIAS SIMPLES DE 30 ADMINISTRADORES DE EAI CADA UNA

Muestra número	Media muestral ($\bar{x}$)	Proporción muestral ($\bar{p}$)
1	51 814	0.63
2	52 670	0.70
3	51 780	0.67
4	51 588	0.53
.	.	.
.	.	.
.	.	.
500	51 752	0.50

de probabilidad. Como los distintos valores que toma $\bar{x}$ son resultado de distintas muestras aleatorias simples, a la distribución de probabilidad de $\bar{x}$ se le conoce como **distribución muestral** de $\bar{x}$. Conocer esta distribución muestral y sus propiedades permitirá hacer declaraciones de probabilidad de qué tan cerca está la media muestral $\bar{x}$ de la media poblacional μ.

De regreso a la figura 7.1, se necesitaría enumerar todas las muestras posibles de 30 administradores y calcular cada una de las medias muestrales para determinar totalmente la distribución muestral de $\bar{x}$. Sin embargo, el histograma de 500 valores $\bar{x}$ da una aproximación a esta distribución muestral. En esta aproximación se observa la apariencia de curva de campana de esta distribución. Además, la mayor concentración de valores de $\bar{x}$ y la media de los 500 valores de $\bar{x}$ se encuentran cerca de la media poblacional $\mu = \$51\,800$. En la sección siguiente se describirán más detalladamente las propiedades de la distribución muestral de $\bar{x}$.

Los 500 valores de las proporciones muestrales $\bar{p}$ se resumen en el histograma de frecuencias relativas de la figura 7.2. Como ocurre con $\bar{x}$, $\bar{p}$ es una variable aleatoria. Si se tomara cada muestra posible de tamaño 30 y para cada muestra se calculara el valor $\bar{p}$, la distribución de probabilidad que se obtuviera sería la distribución muestral de $\bar{p}$. En la figura 7.2, el histograma de frecuencias relativas de los 500 valores muestrales da una idea general de la apariencia de la distribución muestral de $\bar{p}$.

En la práctica sólo se selecciona una muestra aleatoria simple de la población. En esta sección el proceso de muestreo se repitió 500 veces para ilustrar que es posible tomar muchas mues-

TABLA 7.5 DISTRIBUCIÓN DE FRECUENCIAS DE $\bar{x}$ EN 500 MUESTRAS ALEATORIAS SIMPLES DE 30 ADMINISTRADORES DE EAI CADA UNA

Salario anual medio ($)	Frecuencia	Frecuencia relativa
49 500.00–49 999.99	2	0.004
50 000.00–50 499.99	16	0.032
50 500.00–50 999.99	52	0.104
51 000.00–51 499.99	101	0.202
51 500.00–51 999.99	133	0.266
52 000.00–52 499.99	110	0.220
52 500.00–52 999.99	54	0.108
53 000.00–53 499.99	26	0.052
53 500.00–53 999.99	6	0.012
	Totales 500	1.000

FIGURA 7.1 HISTOGRAMA DE LAS FRECUENCIAS RELATIVAS DE LOS VALORES DE $\bar{x}$ OBTENIDOS EN 500 MUESTRAS ALEATORIAS SIMPLES DE 30 ADMINISTRADORES CADA UNA

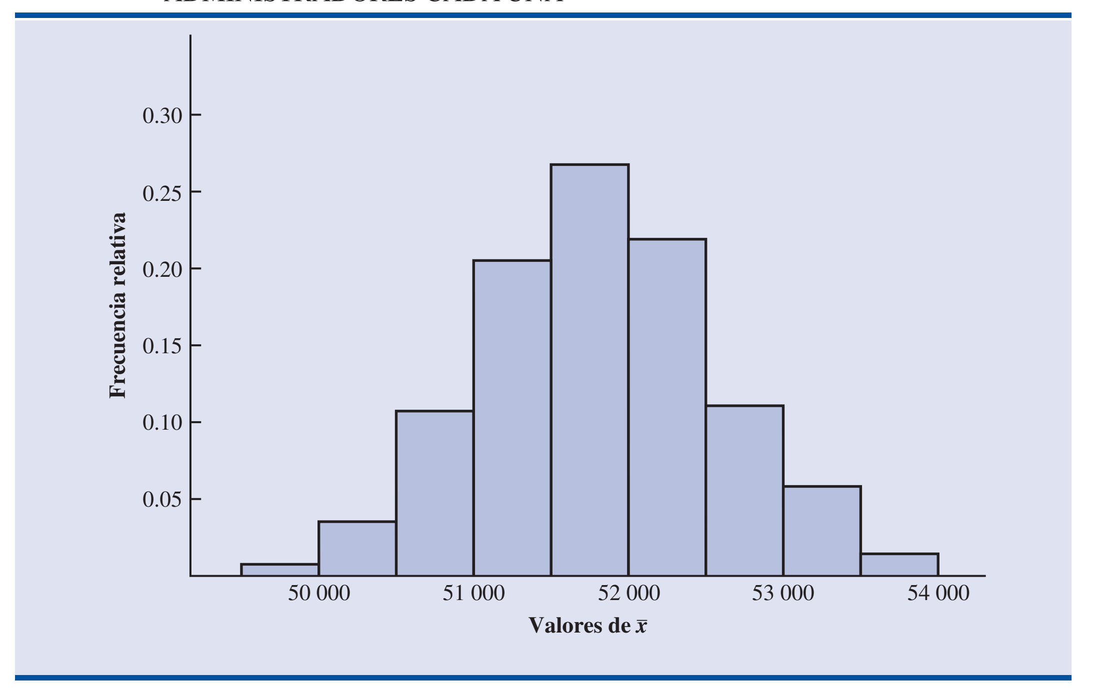
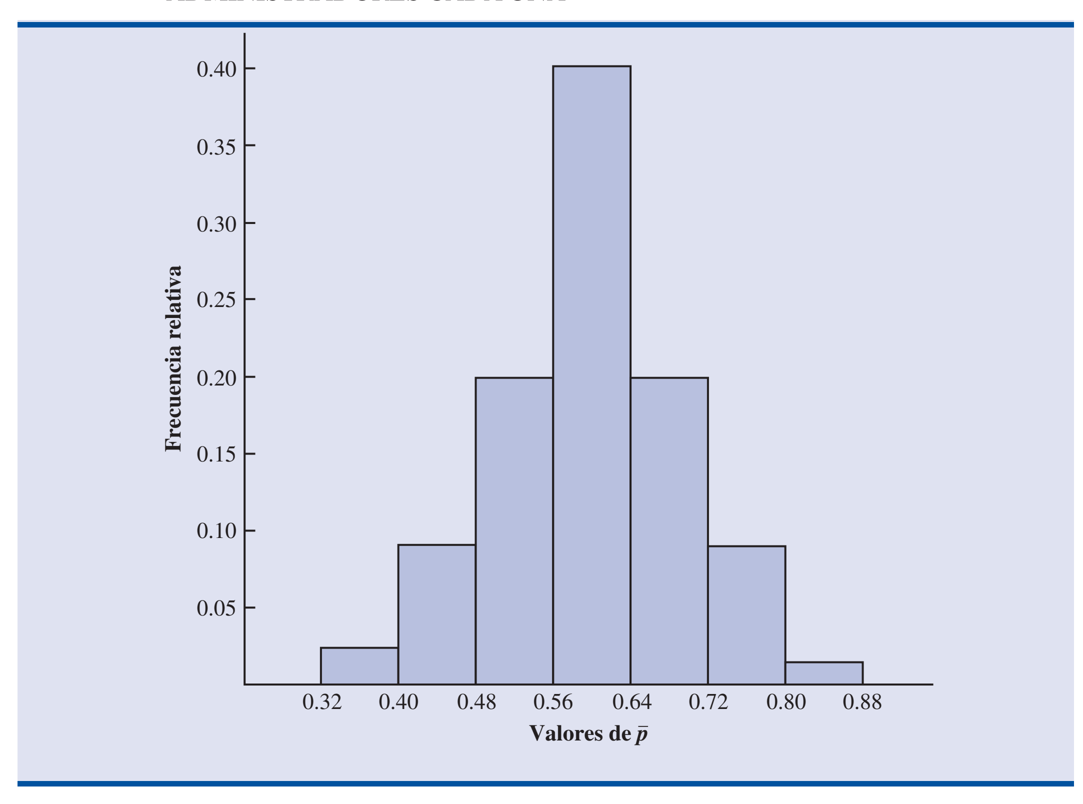


FIGURA 7.2 HISTOGRAMA DE LAS FRECUENCIAS RELATIVAS DE LOS VALORES DE $\bar{p}$ OBTENIDOS EN 500 MUESTRAS ALEATORIAS SIMPLES DE 30 ADMINISTRADORES CADA UNA



tras diferentes y que diferentes muestras darán valores distintos de los estadísticos muestrales $\bar{x}$ y $\bar{p}$. A la distribución muestral de cualquier estadístico determinado se le llama distribución muestral del estadístico. En la sección 7.5 se presentan las características de la distribución muestral de $\bar{x}$. En la sección 7.6 se muestran las características de la distribución muestral de $\bar{p}$.



Distribución muestral de $\bar{x}$

En la sección anterior se dijo que la media muestral $\bar{x}$ es una variable aleatoria y que a su distribución de probabilidad se le llama distribución muestral de $\bar{x}$.

DISTRIBUCIÓN MUESTRAL DE $\bar{x}$

La distribución muestral de $\bar{x}$ es la distribución de probabilidad de todos los valores de la media muestral $\bar{x}$.

En esta sección se describen las propiedades de la distribución muestral de $\bar{x}$. Como ocurre con otras distribuciones de probabilidad estudiadas, la distribución muestral de $\bar{x}$ tiene un valor esperado, una desviación estándar y una forma característica. Para empezar se considerará la media de todos los valores de $\bar{x}$, a la que se conoce como valor esperado de $\bar{x}$.

Valor esperado de $\bar{x}$

En el problema de muestreo de EAI se vio que en distintas muestras aleatorias simples se obtienen valores diferentes para la media muestral $\bar{x}$. Como la variable aleatoria $\bar{x}$ puede tener muchos valores diferentes, suele ser de interés conocer la media de todos los valores de $\bar{x}$ que se obtienen con diferentes muestras aleatorias simples. La media de la variable aleatoria $\bar{x}$ es el valor esperado de $\bar{x}$. Sea $E(\bar{x})$ el valor esperado de $\bar{x}$ y μ la media de la población de la que se selecciona una muestra aleatoria simple. Se puede demostrar que cuando se emplea el muestreo aleatorio simple, $E(\bar{x})$ y μ son iguales.

El valor esperado de $\bar{x}$ es igual a la media de la población de la que se tomó la muestra.

VALOR ESPERADO DE $\bar{x}$

$$E(\bar{x}) = \mu \qquad \textbf{(7.1)}$$

donde

$$E(\bar{x}) = \text{valor esperado de } \bar{x}$$
$$\mu = \text{media poblacional}$$

Esto enseña que usando el muestreo aleatorio simple, el valor esperado o media de la distribución muestral de $\bar{x}$ es igual a la media de la población. En la sección 7.1 se vio que el sueldo anual medio de los administradores de EAI es $\mu = \$51\ 800$. Por tanto, de acuerdo con la ecuación (7.1), la media de todas las medias muestrales en el estudio de EAI es también \$51 800.

Cuando el valor esperado de un estimador puntual es igual al parámetro poblacional, se dice que el estimador puntual es **insesgado**. Por tanto, la ecuación (7.1) muestra que $\bar{x}$ es un estimador insesgado de la media poblacional μ.

Desviación estándar de $\bar{x}$

Ahora se definirá la desviación estándar de la distribución muestral de $\bar{x}$. Se empleará la notación siguiente.

$$\begin{aligned} \sigma_{\bar{x}} &= \text{desviación estándar de } \bar{x} \\ \sigma &= \text{desviación estándar de la población} \\ n &= \text{tamaño de la muestra} \\ N &= \text{tamaño de la población} \end{aligned}$$

Es posible demostrar que usando el muestreo aleatorio simple, la desviación estándar de $\bar{x}$ depende de si la población es finita o infinita. Las dos fórmulas para la desviación estándar son las siguientes.

DESVIACIÓN ESTÁNDAR DE $\bar{x}$

Población finita	*Población infinita*	
$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	**(7.2)**

Al comparar las dos fórmulas se ve que el factor $\sqrt{(N-n)/(N-1)}$ se requiere cuando la población es finita, pero no cuando es infinita. A este factor se le conoce como **factor de corrección para una población finita**. En muchas de las situaciones prácticas de muestreo, se encuentra que aunque la población sea finita, es "grande", mientras que el tamaño de la muestra es "pequeño". En estos casos el factor de corrección para una población finita $\sqrt{(N-n)/(N-1)}$ es casi igual a 1. Por tanto, la diferencia entre el valor de la desviación estándar de $\bar{x}$ en el caso de poblaciones finita o infinitas se vuelve despreciable. Entonces, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ es una buena aproximación a la desviación estándar de $\bar{x}$, aun cuando la población sea finita. Esta observación lleva al siguiente lineamiento, o regla general, para calcular la desviación estándar de $\bar{x}$.

USO DE LA EXPRESIÓN SIGUIENTE PARA CALCULAR LA DESVIACIÓN ESTÁNDAR DE $\bar{x}$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{(7.3)}$$

siempre que

1. La población sea infinita; o
2. La población sea finita y el tamaño de la muestra sea menor o igual a 5% del tamaño de la población; es decir, $n/N \leq 0.05$.

En el problema 21 se muestra que cuando $n/N \leq 0.05$, el factor de corrección para una población finita tiene poco efecto en el valor de $\sigma_{\bar{x}}$.

En los casos en que $n/N > 0.05$, para calcular $s_{\grave{x}}$ deberá usarse la versión para poblaciones finitas de la fórmula (7.2). En este libro, a menos que se indique otra cosa, se supondrá que el tamaño de la población es "grande", $n/N \leq 0.05$, y se usará la expresión (7.3) para calcular $\sigma_{\bar{x}}$.

El término error estándar se usa en la inferencia estadística para referirse a la desviación estándar de un estimador puntual.

Para calcular $\sigma_{\bar{x}}$ se necesita conocer σ, la desviación estándar de la población. Para subrayar, aún más, la diferencia entre $\sigma_{\bar{x}}$ y σ, a la desviación estándar de $\bar{x}$, $\sigma_{\bar{x}}$ se le llama **error estándar** de la media. En general, el término *error estándar* se refiere a la desviación estándar de un estimador puntual. Más adelante se verá que el valor del error estándar de la media ayuda a determinar qué tan lejos puede estar la media muestral de la media poblacional. Ahora, de nuevo con el ejemplo de EAI se calcula el error estándar de la media correspondiente a las muestras aleatorias simples de 30 administradores de EAI.

En la sección 7.1 se halló que la desviación estándar de los sueldos anuales en la población de los 2500 administradores de EAI era $\sigma = 4000$. En este caso la población es finita, $N = 2500$. Sin embargo, como el tamaño de la muestra es 30, se tiene $n/N = 30/2500 = 0.012$. Como el tamaño de la muestra es menor que 5% del tamaño de la población, se puede ignorar el factor de corrección para una población finita y usar la ecuación (7.3) para calcular el error estándar.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{30}} = 730.3$$

Forma de la distribución muestral de $\bar{x}$

Los resultados anteriores respecto al valor esperado y a la desviación estándar en la distribución muestral de $\bar{x}$ son aplicables a cualquier población. El paso final en la identificación de las características de la distribución muestral de $\bar{x}$ es determinar la forma de la distribución muestral. Se considerarán dos casos: 1. La población tiene distribución normal, y 2. La población no tiene distribución normal.

La población tiene distribución normal. En muchas situaciones es razonable suponer que la población de la que se seleccionó la muestra aleatoria simple tenga distribución normal o casi normal. Cuando la población tiene distribución normal, la distribución muestral de $\bar{x}$ está distribuida normalmente sea cual sea el tamaño de la muestra.

La población no tiene distribución normal. Cuando la población de la que se tomó la muestra aleatoria simple no tiene distribución normal, el **teorema del límite central** ayuda a determinar la forma de la distribución muestral de $\bar{x}$. El enunciado del teorema del límite central aplicado a la distribución muestral de $\bar{x}$ dice lo siguiente.

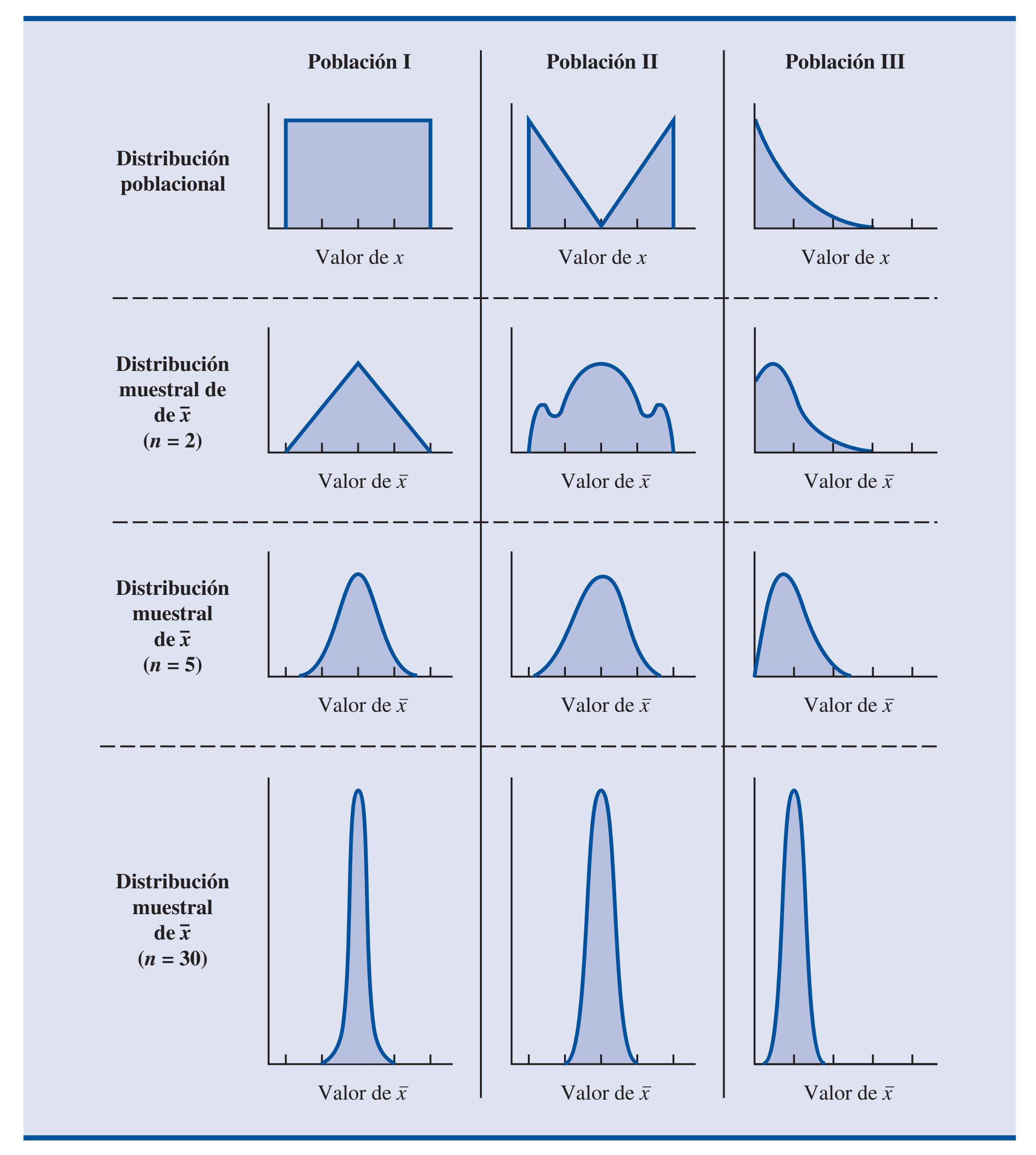
TEOREMA DEL LÍMITE CENTRAL

Cuando se seleccionan muestras aleatorias simples de tamaño n de una población, la distribución muestral de la media muestral $\bar{x}$ puede aproximarse mediante una distribución normal a medida que el tamaño de la muestra se hace grande.

En la figura 7.3 se muestra cómo funciona el teorema del límite central en tres poblaciones diferentes; cada columna se refiere a una de las poblaciones. En el primer renglón de la figura se muestra que ninguna de las tres poblaciones está distribuida normalmente. La población I tiene una distribución uniforme. A la población II se le conoce como distribución en forma de orejas de conejo. Esta distribución es simétrica, pero los valores más probables se encuentran en las colas de la distribución. La forma de la población III se parece a una distribución exponencial; es sesgada a la derecha.

En los tres renglones siguientes de la figura 7.3 se muestran las formas de las distribuciones muestrales para tamaños de muestras $n = 2$, $n = 5$ y $n = 30$. Cuando el tamaño de la muestra es 2, se observa que cada distribución muestral tiene una forma diferente a la distribución poblacional correspondiente. Con muestras de tamaño 5, se observa que las formas de las distribuciones muestrales en los casos de las poblaciones I y II empiezan a parecerse a la forma de una distribución normal. En el caso de la población III, aun cuando la forma de la distribución muestral empieza a ser parecida a una distribución normal, todavía se observa cierto sesgo a la derecha.

FIGURA 7.3 ILUSTRACIÓN DEL TEOREMA DEL LÍMITE CENTRAL CON TRES POBLACIONES



Por último, para muestras de tamaño 30, las formas de cada una de las tres distribuciones muestrales es aproximadamente normal.

Desde el punto de vista de la práctica, será conveniente saber qué tan grande necesita ser el tamaño de la muestra para que aplique el teorema del límite central y pueda suponer que la forma de la distribución muestral es aproximadamente normal. En las investigaciones estadísticas se ha estudiado este problema en distribuciones muestrales de $\bar{x}$ de muy diversas poblaciones y para muy diversos tamaños de muestras. Lo que se acostumbra hacer en la práctica es suponer que, en la mayor parte de las aplicaciones, la distribución muestral de $\bar{x}$ se puede aproximar mediante una distribución normal siempre que la muestra sea de tamaño 30 o mayor. En los casos en que la población es muy sesgada o existen observaciones atípicas, pueden necesitarse muestras de tamaño 50. Por último, si la población es discreta, el tamaño de muestra necesario para la aproximación normal suele depender de la proporción poblacional. Más acerca de este tema se dirá cuando se estudie la distribución muestral de $\bar{p}$ en la sección 7.6.

Distribución muestral de $\bar{x}$ en el problema EAI

En el problema de EAI, para el que ya previamente se mostró que $E(\bar{x}) = \$51\,800$ y $\sigma_{\bar{x}} = 730.3$, no se cuenta con ninguna información acerca de la distribución de la población; puede estar o no distribuida normalmente. Si la población tiene una distribución normal, la distribución muestral de $\bar{x}$ estará distribuida normalmente. Si la población no tiene una distribución normal, la muestra aleatoria simple de 30 administradores y el teorema del límite central permiten concluir que la distribución muestral de $\bar{x}$ puede aproximarse mediante una distribución normal. En cualquiera de los casos, se concluye que la distribución muestral de $\bar{x}$ se describe mediante una distribución normal como la que se muestra en la figura 7.4.

Valor práctico de la distribución muestral de $\bar{x}$

Siempre que se seleccione una muestra aleatoria simple y se use el valor de la media muestral para estimar el valor de la media poblacional μ, no se podrá esperar que la media muestral sea exactamente igual a la media poblacional. La razón práctica por la que interesa la distribución muestral de $\bar{x}$ es que se puede usar para proporcionar información probabilística acerca de la diferencia entre la media muestral y la media poblacional. Para demostrar este uso, se retomará el problema de EAI.

Suponga que el director de personal cree que la media muestral será una estimación aceptable de la media poblacional si la primera está a más o menos de \$500 de la media poblacional. Sin embargo, no es posible garantizar que la media muestral esté a más o menos de \$500 de la media poblacional. En efecto, en la tabla 7.5 y en la figura 7.1 se observa que algunas de las 500 medias muestrales difieren en más de \$2000 de la media poblacional. Entonces hay que pensar en el requerimiento del director de personal en términos de probabilidad. Es decir, al director de personal le interesa la interrogante siguiente: "¿Cuál es la probabilidad de que la media muestral obtenida usando una muestra aleatoria simple de 30 administradores de EAI, se encuentre a más o menos de \$500 de la media poblacional?"

Como ya se han identificado las propiedades de la distribución muestral de $\bar{x}$ (véase figura 7.4), se usará esta distribución para contestar esta interrogante probabilística. Observe la distribución muestral de $\bar{x}$ que se muestra nuevamente en la figura 7.5. Como la media poblacional es \$51 800, el director de personal desea saber cuál es la probabilidad de que $\bar{x}$ esté entre \$51 300 y \$52 300. Esta probabilidad corresponde al área sombreada de la distribución muestral que apa-


FIGURA 7.4 DISTRIBUCIÓN MUESTRAL DE $\bar{x}$ PARA EL SUELDO ANUAL EN UNA MUESTRA ALEATORIA SIMPLE DE 30 ADMINISTRADORES

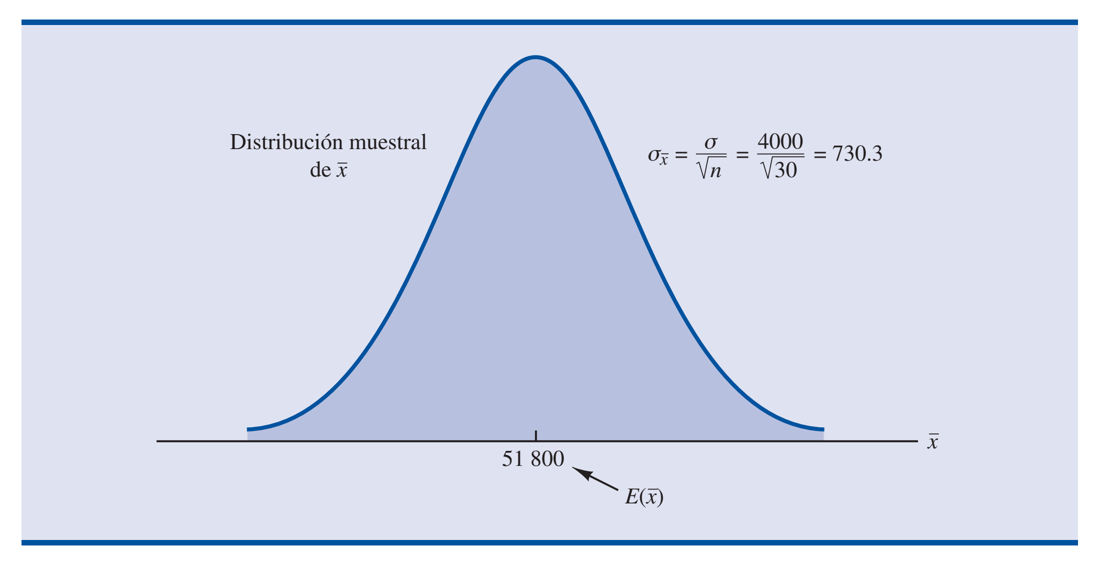
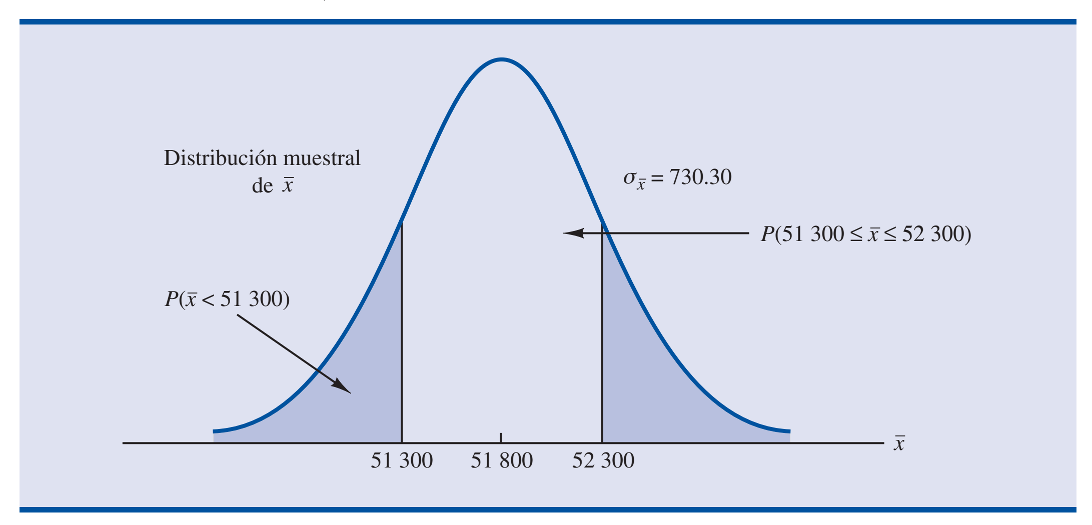


FIGURA 7.5 PROBABILIDAD DE QUE UNA MEDIA MUESTRAL DE UNA MUESTRA ALEATORIA SIMPLE DE 30 ADMINISTRADORES DE EAI SE ENCUENTRE ENTRE LOS $500 DE LA MEDIA POBLACIONAL



rece en la figura 7.5. Como la distribución muestral está distribuida normalmente y su media es $51 800 y el error estándar de la media es 730.3, se usa la tabla de probabilidad normal estándar para hallar el área o probabilidad.

Primero se calcula el valor de z en el extremo superior de este intervalo (52 300) y se usa la tabla para hallar el área bajo la curva a la izquierda de ese punto (área hacia la cola izquierda). Después se calcula el valor de z en el extremo inferior de este intervalo (51 300) y se usa la tabla para hallar el área bajo la curva a la izquierda de este punto (otra área hacia la cola izquierda). Al restar la segunda área de la primera, se obtiene la probabilidad buscada.

En $\bar{x} = 52\,300$, se tiene

$$z = \frac{52\,300 - 51\,800}{730.30} = 0.68$$

En la tabla de probabilidad normal estándar la probabilidad acumulada (área a la izquierda de $z = 0.68$) es 0.7517.

En $\bar{x} = 51\,300$, se tiene

$$z = \frac{51\,300 - 51\,800}{730.30} = -0.68$$

El área bajo la curva a la izquierda de $z = -0.68$ es 0.2483. Por tanto, $P(51\,300 \le \bar{x} \le 52\,300) = P(z \le 0.68) - P(z < -0.68) = 0.7517 - 0.2483 = 0.5034$.

La distribución muestral de $\bar{x}$ se usa para obtener información probabilística acerca de qué tan cerca se encuentra la media muestral $\bar{x}$ de la media poblacional μ.

Estos cálculos indican que hay una probabilidad de 0.5034 de que con una muestra aleatoria simple de 30 administradores de EAI se obtenga una media muestral $\bar{x}$ que esté a más o menos de $500 de la media poblacional. Por tanto, la probabilidad de que la diferencia entre $\bar{x}$ y $\mu = \$51{,}800$ sea superior a $500 es $1 - 0.5034 = 0.4966$. En otras palabras, una muestra aleatoria simple de 30 administradores de EAI tiene aproximadamente 50/50 oportunidades de tener una media muestral que no difiera de la media poblacional en más de los aceptables $500. Quizá deba pensarse en una muestra de tamaño mayor. Se explorará esta posibilidad considerando la relación entre el tamaño de la muestra y la distribución muestral de $\bar{x}$.

Relación entre el tamaño de la muestra y la distribución muestral de $\bar{x}$

Suponga que en el problema de muestreo de EAI se toma una muestra aleatoria simple de 100 administradores en lugar de los 30 considerados. La intuición indica que teniendo más datos proporcionados por una muestra mayor, la media muestral basada en $n = 100$ proporcionará una mejor estimación de la media poblacional que una media muestral basada en $n = 30$. Para ver cuánto es mejor, se considerará la relación entre el tamaño de la muestra y la distribución muestral de $\bar{x}$.

Primero observe que $E(\bar{x}) = \mu$ independientemente del tamaño de la muestra. Entonces, la media de todos los valores posibles de $\bar{x}$ es igual a la media poblacional μ independientemente del tamaño n de la muestra. Pero, el error estándar de la media, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, está relacionado con la raíz cuadrada del tamaño de la muestra. Siempre que el tamaño de la muestra aumente, el error estándar de la media $\sigma_{\bar{x}}$ disminuirá. Con $n = 30$, el error estándar de la media en el problema de EAI es 730.3. Sin embargo, aumentando el tamaño de la muestra $n = 100$, el error estándar de la media disminuye a

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{100}} = 400$$

En la figura 7.6 se muestran las distribuciones muestrales de $\bar{x}$ correspondientes a $n = 30$ y a $n = 100$. Como la distribución muestral con $n = 100$ tiene un error estándar más pequeño, habrá menos variación entre los valores de $\bar{x}$ y éstos tenderán a estar más cerca de la media poblacional que los valores de $\bar{x}$ con $n = 30$.

La distribución muestral de $\bar{x}$, en el caso $n = 100$, puede emplearse para calcular la probabilidad de que una muestra aleatoria simple de 100 administradores de EAI dé una media muestral que no difiera de los \$500 de la media poblacional. Como la distribución muestral es normal y su media es \$51 800 y el error estándar de la media es 400, se emplea la tabla de probabilidad normal estándar para hallar el área o la probabilidad.

Para $\bar{x} = 52\,300$ (véase figura 7.7) se tiene

$$z = \frac{52\,300 - 51\,800}{400} = 1.25$$

FIGURA 7.6 COMPARACIÓN ENTRE LAS DISTRIBUCIONES MUESTRALES DE` $\bar{x}$ CON MUESTRAS ALEATORIAS SIMPLES DE TAMAÑO $n = 30$ ADMINISTRADORES DE EAI Y CON MUESTRAS DE TAMAÑO $n = 100$ ADMINISTRADORES

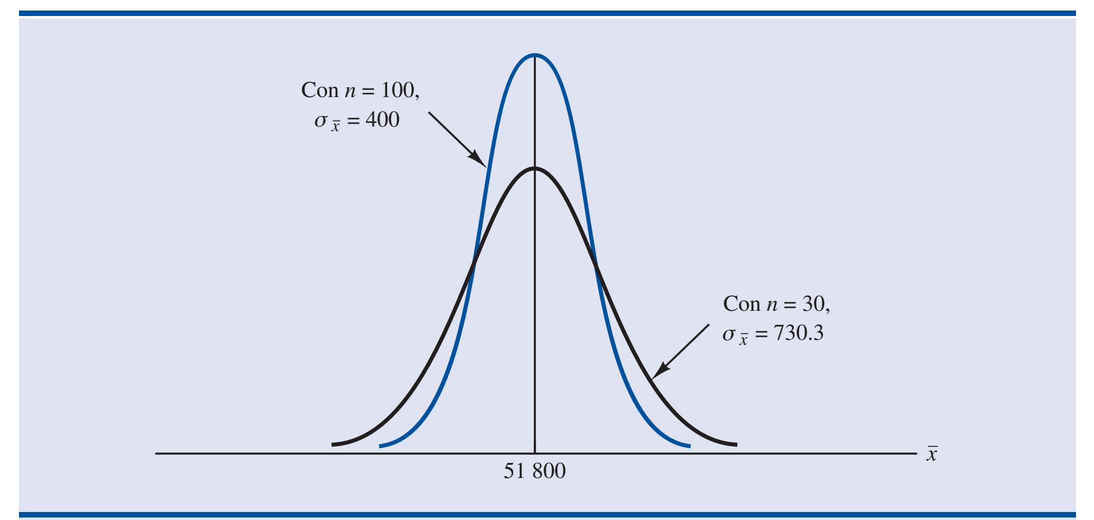
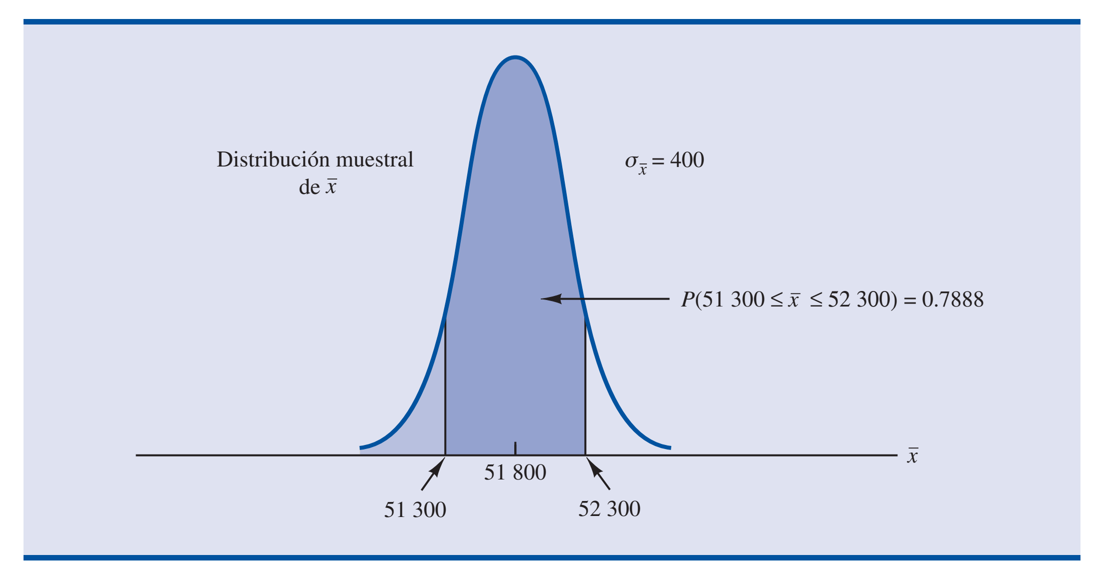


FIGURA 7.7 PROBABILIDAD DE QUE LA MEDIA MUESTRAL NO DIFIERA EN MÁS DE $500 DE LA MEDIA POBLACIONAL USANDO UNA MUESTRA ALEATORIA SIMPLE DE 100 ADMINISTRADORES DE EAI



En la tabla de probabilidad normal estándar se encuentra que la probabilidad acumulada correspondiente a $z = 1.25$ es 0.8944.

Para $\bar{x} = 51\,300$, se tiene

$$z = \frac{51\,300 - 51\,800}{400} = -1.25$$

La probabilidad acumulada correspondiente a $z = -1.25$ es 0.1056. Por tanto, $P(51\,300 \leq \bar{x} \leq 52\,300) = P(z \leq 1.25) - P(z \leq -1.25) = 0.8944 - 0.1056 = 0.7888$. Entonces, aumentando el tamaño de la muestra de 30 a 100 administradores de EAI, la probabilidad de obtener una muestra aleatoria simple que esté entre los $500 de la media poblacional aumenta de 0.5034 a 0.7888.

Aquí, el punto importante es que cuando aumenta el tamaño de la muestra, el error estándar de la media disminuye. Así, una muestra de mayor tamaño proporciona mayor probabilidad de que la media muestral esté dentro de una distancia determinada de la media poblacional.

NOTAS Y COMENTARIOS

1. Al presentar la distribución muestral de $\bar{x}$ para el problema de EAI, se aprovechó que se conocían la media poblacional $\mu = 51\,800$ y la desviación estándar poblacional $\sigma = 4000$. Sin embargo, lo usual es que los valores de la media poblacional μ y de la desviación estándar poblacional σ, que se necesitan para determinar la distribución muestral de $\bar{x}$, no se conozcan. En el capítulo 8 se verá cómo se usan la media muestral $\bar{x}$ y la desviación estándar muestral s cuando no se conocen μ y σ.

2. La demostración del teorema del límite central requiere observaciones independientes en la muestra. Esta condición se satisface cuando se trata de poblaciones infinitas y cuando se trata de poblaciones finitas, si el muestreo se hace con reemplazo. Aunque el teorema del límite central no se refiere directamente a muestreos sin reemplazo de poblaciones finitas, se aplican los hallazgos del teorema del límite central cuando la población es de tamaño grande.

Ejercicios

Métodos

18. La media de una población es 200 y su desviación estándar es 50. Se va a tomar una muestra aleatoria simple de tamaño 100 y se usará la media muestral para estimar la media poblacional.
    a. ¿Cuál es el valor esperado de $\bar{x}$?
    b. ¿Cuál es la desviación estándar de $\bar{x}$?
    c. Muestre la distribución muestral de $\bar{x}$
    d. ¿Qué muestra la distribución muestral de $\bar{x}$?



19. La media de una población es 200 y su desviación estándar es 50. Suponga que se selecciona una variable aleatoria simple de tamaño 100 y se usa $\bar{x}$ para estimar μ
    a. ¿Cuál es la probabilidad de que la diferencia entre la media muestral y la media poblacional no sea mayor que ± 5?
    b. ¿De que la diferencia entre la media muestral y la media poblacional no sea mayor que ± 10?
20. Suponga que la desviación estándar poblacional es $\sigma = 25$. Calcule el error estándar de la media, $\sigma_{\bar{x}}$, con muestras de tamaño 50, 100, 150 y 200. ¿Qué puede decir acerca del error estándar de la media conforme el tamaño de la muestra aumenta?
21. Suponga que de una población en la que $\sigma = 10$ se toma una muestra aleatoria simple de tamaño 50. Halle el valor del error estándar de la media en cada uno de los casos siguientes (si es necesario use el factor de corrección para una población finita).
    a. El tamaño de la población es infinito.
    b. El tamaño de la población es $N = 50\,000$.
    c. El tamaño de la población es $N = 5000$.
    d. El tamaño de la población es $N = 500$.

Aplicaciones

22. Regrese al problema de los administradores de EAI. Suponga que se usa una muestra aleatoria simple de 60 administradores.
    a. Dibuje la distribución muestral de $\bar{x}$ si se emplean muestras aleatorias simples de tamaño 60.
    b. ¿Qué pasa con la distribución muestral de $\bar{x}$ si se usan muestras aleatorias simples de tamaño 120?
    c. ¿Qué puede decir acerca de lo que le pasa a la distribución muestral de $\bar{x}$ conforme el tamaño de la muestra aumenta? ¿Parece ser lógica esta generalización? Explique.



23. En el problema de EAI (véase figura 7.5), se mostró que con $n = 30$, la probabilidad de que la media muestral no difiriera más de \$500 de la media poblacional era 0.5034.
    a. ¿Cuál es la probabilidad de que la media muestral no difiera más de \$500 de la media poblacional si se usa una muestra de tamaño 60?
    b. Responda el inciso a si el tamaño de la muestra es 120.
24. El costo medio de la colegiatura en una universidad estatal de Estados Unidos es \$4260 anuales. Considere este valor como media poblacional y asuma que la desviación estándar poblacional es $\sigma = \$900$. Suponga que selecciona una muestra aleatoria de 50 universidades.
    a. Presente la distribución muestral de $\bar{x}$ como media muestral de la colegiatura en las 50 universidades.
    b. ¿Cuál es la probabilidad de que la muestra aleatoria simple proporcione una media muestral que no difiera de la media poblacional en más de \$250?
    c. ¿Cuál es la probabilidad de que la muestra aleatoria simple proporcione una media muestral que no difiera de la media poblacional en más de \$100?
25. El College Board American College Testing Program informa que en el examen de admisión a las universidades, a nivel nacional, la media poblacional de las puntuaciones que se obtienen es $\mu = 1020$ (*The World Almanac 2003*). Suponga que la desviación estándar poblacional es $\sigma = 100$.

a. ¿Cuál es la probabilidad de que en una muestra aleatoria de 75 estudiantes la media muestral de las puntuaciones no difiera más de 10 puntos de la media poblacional?
b. ¿Cuál es la probabilidad de que en una muestra aleatoria de 75 estudiantes la media muestral de las puntuaciones no difiera más de 20 puntos de la media poblacional?

26. El costo medio anual de un seguro para automóvil es de $939 (*CNBC*, 23 de febrero de 2006). Suponga que la desviación estándar es $\sigma = \$245$.
    a. ¿Cuál es la probabilidad de que una muestra aleatoria simple de pólizas de seguros de automóvil la media muestral no difiera más de $25 de la media poblacional si el tamaño de la muestra es 30, 50, 100 y 400?
    b. ¿Qué ventaja tiene una muestra grande cuando se quiere estimar la media poblacional?

27. *BusinessWeek* realizó una encuesta entre los estudiantes que terminaban sus estudios en los 30 programas de una maestría (*BusinessWeek*, 22 de septiembre de 2003). De acuerdo con esta encuesta el salario medio anual de una mujer y de un hombre 10 años después de terminar sus estudios es $117 000 y $168 000, respectivamente. Suponga que la desviación estándar entre los salarios de las mujeres es $25 000 y entre los salarios de los hombres es $40 000.
    a. ¿Cuál es la probabilidad de que en una muestra aleatoria simple de 40 hombres la media muestral no difiera más de $10 000 de la media poblacional de $168 000?
    b. ¿Cuál es la probabilidad de que en una muestra aleatoria simple de 40 mujeres la media muestral no difiera más de $10 000 de la media poblacional de $117 000?
    c. ¿En cuál de los dos casos, inciso a o inciso b, hay más probabilidad de obtener una media muestral que no difiera en más de $10 000 de la media poblacional? ¿Por qué?
    d. ¿Cuál es la probabilidad de que en una muestra aleatoria simple de 100 hombres, la media muestral no difiera en más de $4000 de la media poblacional?

28. Un hombre golfista tiene una puntuación promedio de 95 y una mujer de 106 (*Golf Digest*, abril de 2006). Considere estos valores como medias poblacionales de los hombres y de las mujeres y suponga que la desviación estándar poblacional es $\sigma = 14$ golpes en ambos casos. Se tomará una muestra aleatoria simple de 40 golfistas hombres y otra de 45 mujeres golfistas
    a. Dé la distribución muestral de $\bar{x}$ correspondiente a los hombres golfistas.
    b. ¿Cuál es la probabilidad de que, en el caso de los hombres golfistas, la media muestral no difiera en más de 3 golpes de la media poblacional?
    c. ¿Cuál es la probabilidad de que, en el caso de las mujeres golfistas, la media muestral no difiera en más de 3 golpes de la media poblacional?
    d. ¿En cuál de los casos, inciso a o inciso b, es mayor la probabilidad de que la media muestral no difiera en más de 3 golpes de la media poblacional? ¿Por qué?

29. En el norte de Kentucky (*The Cincinnati Enquirer*, 21 de enero de 2006) el precio promedio de la gasolina sin plomo era $2.34. Use este precio como media poblacional y suponga que la desviación estándar poblacional es $0.20.
    a. ¿Cuál es la probabilidad de que el precio medio en una muestra de 30 gasolineras no difiera en más de $0.30 de la media poblacional?
    b. ¿Cuál es la probabilidad de que el precio medio en una muestra de 50 gasolineras no difiera en más de $0.30 de la media poblacional?
    c. ¿Cuál es la probabilidad de que el precio medio en una muestra de 100 gasolineras no difiera en más de $0.30 de la media poblacional?
    d. ¿Recomendaría usted alguno de los tamaños muestrales de los incisos a, b o c para que la probabilidad de que el precio muestral no difiriera en más de $0.30 de la media muestral fuera $0.95?

30. Para estimar la edad media de una población de 4000 empleados se toma una muestra de 40 empleados.
    a. ¿Usted usaría el factor de corrección para una población finita en el cálculo del error estándar de la media? Explique.

b. Si la desviación estándar poblacional es $\sigma = 8.2$ años, calcule el error estándar con y sin el factor de corrección para una población finita. ¿Cuál es la base para ignorar el factor de corrección para la población finita, si $n/N \leq 0.05$?
c. ¿Cuál es la probabilidad de que la media muestral de las edades de los empleados no difiera en más de ± 2 años de la media poblacional de las edades?



7.6 Distribución muestral de $\bar{p}$

La proporción muestral $\bar{p}$ es el estimador puntual de la proporción poblacional p. La fórmula para calcular la proporción muestral es

$$\bar{p} = \frac{x}{n}$$

donde

x = número de elementos de la muestra que poseen la característica de interés
n = tamaño de la muestra

Como se indicó en la sección 7.4, la proporción muestral $\bar{p}$ es una variable aleatoria y su distribución de probabilidad se conoce como distribución muestral de $\bar{p}$.

> DISTRIBUCIÓN MUESTRAL DE $\bar{p}$
>
> La distribución muestral de $\bar{p}$ es la distribución de probabilidad de todos los posibles valores de la proporción muestral $\bar{p}$.

Para determinar qué tan cerca está la proporción muestral $\bar{p}$ de la proporción poblacional p, se necesita entender las propiedades de la distribución muestral de $\bar{p}$: el valor esperado de $\bar{p}$, la desviación estándar de $\bar{p}$ y la forma de la distribución muestral de $\bar{p}$.

Valor esperado de $\bar{p}$

El valor esperado de $\bar{p}$, la media de todos los posibles valores de $\bar{p}$, es igual a la proporción poblacional p.

> VALOR ESPERADO DE $\bar{p}$
>
> $$E(\bar{p}) = p \tag{7.4}$$
>
> donde
>
> $E(\bar{p})$ = valor esperado de $\bar{p}$
> p = proporción poblacional

Como $E(\bar{p}) = p$, $\bar{p}$ es un estimador insesgado de p. Recuerde que en la sección 7.1 se encontró que en la población de EAI $p = 0.60$, siendo p la proporción de la población de administradores que han participado en el programa de capacitación de la empresa. Por tanto, el valor esperado de $\bar{p}$ en el problema de muestreo de EAI es 0.60.

Desviación estándar de $\bar{p}$

Como en el caso de la desviación estándar de $\bar{x}$ la desviación estándar de $\bar{p}$ obedece a si la población es finita o infinita. Las dos fórmulas para calcular la desviación estándar de $\bar{p}$ se presentan a continuación.

DESVIACIÓN ESTÁNDAR DE $\bar{p}$

Población finita | *Población infinita*

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad \textbf{(7.5)}$$

Al comparar las dos fórmulas (7.5) se aprecia que la única diferencia es el uso del factor de corrección para una población finita $\sqrt{(N-n)/(N-1)}$.

Como en el caso de la media poblacional $\bar{x}$, la diferencia entre las expresiones para una población finita y para una infinita es despreciable si el tamaño de la población finita es grande en comparación con el tamaño de la muestra. Se seguirá la misma regla recomendada para la media poblacional. Es decir, si la población es finita y $n/N \leq 0.05$ se usará $\sigma_{\bar{p}} = \sqrt{p(1-p)/n}$. Pero, si la población es finita y $n/N > 0.05$, entonces deberá usar el factor de corrección para una población finita. También, a menos que se especifique otra cosa, en este libro se supondrá que el tamaño de la población es grande en comparación al tamaño de la muestra y por tanto, el factor de corrección para una población finita no será necesario.

En la sección 7.5 se usó el error estándar de la media para referirse a la desviación estándar de $\bar{x}$. Se dijo que en general el término error estándar se refiere a la desviación estándar de un estimador puntual. Así, en el caso de proporciones se usa *el error estándar de la proporción* para referirse a la desviación estándar de $\bar{p}$. Ahora se vuelve al ejemplo de EAI para calcular el error estándar de la proporción en la muestra aleatoria simple de los 30 administradores de EAI.

En el estudio de EAI se sabe que la proporción poblacional de administradores que han participado en el programa de capacitación es $p = 0.60$. Como $n/N = 30/2\,500 = 0.012$ se puede ignorar el factor de corrección para una población finita al calcular el error estándar de la proporción. En la muestra aleatoria simple de 30 administradores, $\sigma_{\bar{p}}$ es

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.60(1-0.60)}{30}} = 0.0894$$

Forma de la distribución muestral de $\bar{p}$

Ahora que se conoce la media y la desviación estándar de la distribución muestral de $\bar{p}$, el último paso es determinar la forma de la distribución muestral. La proporción muestral es $\bar{p} = x/n$. En una muestra aleatoria simple de una población grande, el valor de x es una variable aleatoria binomial que indica el número de los elementos de la muestra que tienen la característica de interés. Como n es una constante, la probabilidad de x/n es la misma que la probabilidad de x, lo cual significa que la distribución muestral de $\bar{p}$ también es una distribución de probabilidad discreta y que la probabilidad de cada x/n es la misma que la probabilidad de x.

En el capítulo 6 se mostró que una distribución binomial se aproxima mediante una distribución normal siempre que el tamaño de la muestra sea lo suficientemente grande para satisfacer las dos condiciones siguientes.

$$np \geq 5 \quad \text{y} \quad n(1 - p) \geq 5$$

Suponiendo que se satisfagan estas dos condiciones, la distribución de probabilidad de x en la proporción muestral, $\bar{p} = x/n$, puede aproximarse por medio de una distribución normal. Y como n es una constante, la distribución muestral de $\bar{p}$ también se aproxima mediante una distribución normal. Esta aproximación se formula como sigue:

> La distribución muestral de $\bar{p}$ se aproxima mediante una distribución normal siempre que $np \geq 5$ y $n(1 - p) \geq 5$.

En las aplicaciones prácticas, cuando se requiere una estimación de la proporción poblacional, casi siempre se encuentra que el tamaño de la muestra es suficientemente grande para poder usar la aproximación normal para la distribución muestral de $\bar{p}$.

Recuerde que en el problema de muestreo de EAI la proporción poblacional de administradores que han participado en el programa de capacitación es $p = 0.60$. Con una muestra aleatoria simple de tamaño 30, se tiene $np = 30(0.60) = 18$ y $n(1 - p) = 30(0.40) = 12$. Por tanto, la distribución muestral de $\bar{p}$ se calcula mediante la distribución normal que se muestra en la figura 7.8.

Valor práctico de la distribución muestral de $\bar{p}$

El valor práctico de la distribución muestral de $\bar{p}$ es que permite obtener información probabilística acerca de la diferencia entre la proporción muestral y la proporción poblacional. Por ejemplo, en el problema de EAI, el director de personal desea saber cuál es la probabilidad de obtener un valor de $\bar{p}$ que no difiera en más de 0.05 de la proporción poblacional de los administradores de EAI que han participado en el programa de capacitación. Es decir, ¿cuál es la probabilidad de tener una muestra en la que $\bar{p}$ esté entre 0.55 y 0.65? El área sombreada de la figura 7.9 corres-

FIGURA 7.8 DISTRIBUCIÓN MUESTRAL DE $\bar{p}$, PROPORCIÓN DE ADMINISTRADORES QUE HAN PARTICIPADO EN EL PROGRAMA DE CAPACITACIÓN DE EAI

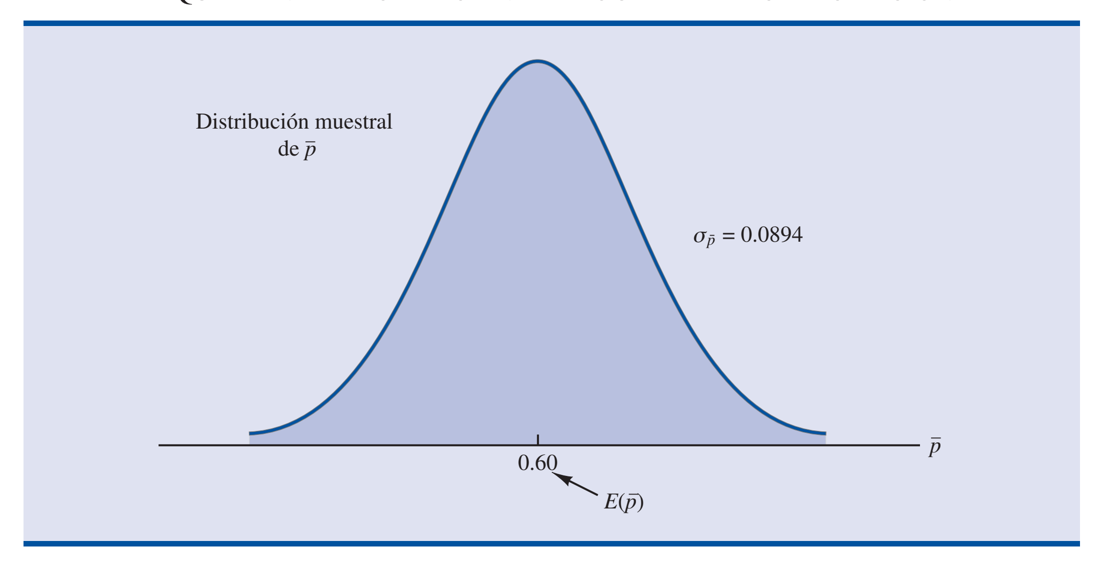
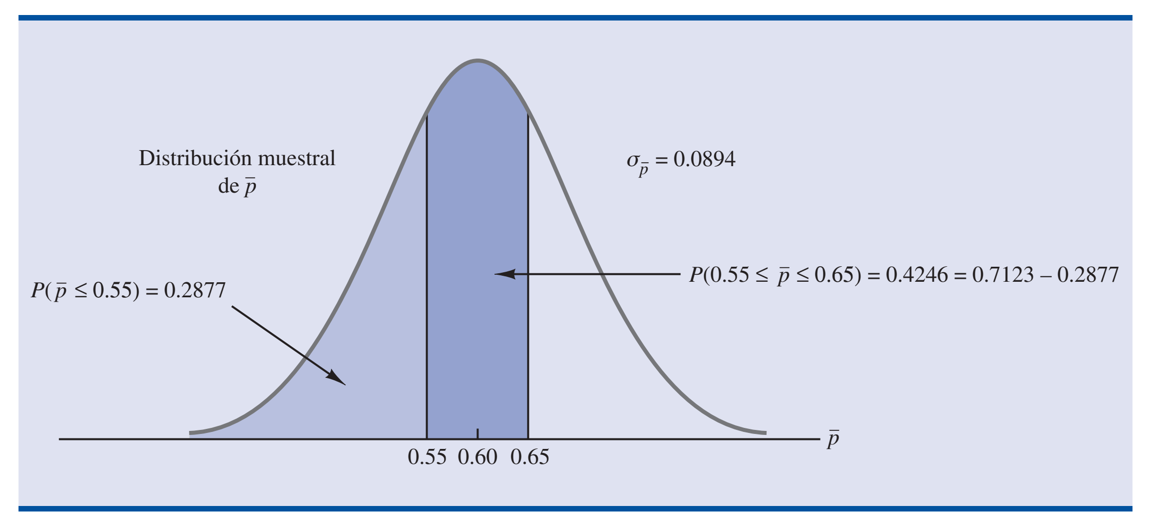


FIGURA 7.9 PROBABILIDAD DE QUE $\bar{p}$ ESTÉ ENTRE 0.55 Y 0.65



ponde a esta probabilidad. A partir de que la distribución muestral de $\bar{p}$ se aproxima mediante una distribución normal con media 0.60 y error estándar de la proporción $\sigma_{\bar{p}} = 0.0894$, se encuentra que la variable aleatoria normal estándar correspondiente a $\bar{p} = 0.65$ tiene el valor $z = (0.65 - 0.60)/0.0894 = 0.56$. En la tabla de probabilidad normal estándar aparece que la probabilidad acumulada que corresponde a $z = 0.56$ es 0.7123. De manera similar para $\bar{p} = 0.55$, se encuentra que $z = (0.55 - 0.60)/0.0894 = -0.56$. En la misma tabla y correspondiente a $z = -0.56$ es 0.2877. De esta manera, la probabilidad de seleccionar una muestra en la cual el valor de $\bar{p}$ no difiera más de 0.05 de la proporción poblacional p está dada por $0.7123 - 0.2877 = 0.4246$.

Si se aumenta el tamaño de la muestra a $n = 100$, el error estándar de la proporción se convierte en

$$\sigma_{\bar{p}} = \sqrt{\frac{0.60(1 - 0.60)}{100}} = 0.049$$

Con una muestra de 100 administradores de EAI, se calcula ahora la probabilidad de que la proporción muestral tenga un valor que no difiera en más de 0.05 de la proporción poblacional. Como la distribución muestral es aproximadamente normal, con media 0.60 y desviación estándar 0.049, se puede usar la tabla de probabilidad normal estándar para hallar el área o probabilidad. Para $\bar{p} = 0.65$, se tiene $z = (0.65 - 0.60)/0.049 = 1.02$. La tabla de probabilidad normal estándar arroja que la probabilidad acumulada correspondiente a $z = 1.02$ es 0.8461. De manera similar, para $\bar{p} = 0.55$, se tiene que $z = (0.55 - 0.60)/0.049 = -1.02$. Se encuentra que la probabilidad acumulada correspondiente a $z = -1.02$ es 0.1539. Por tanto, si el tamaño de la muestra aumenta de 30 a 100, la probabilidad de que la proporción muestral $\bar{p}$ no difiera en más de 0.05 de la proporción poblacional aumenta a $0.8461 - 0.1539 = 0.6922$.

Ejercicios

Métodos

31. De una muestra aleatoria de tamaño 100 de una población en la que $p = 0.40$.
    a. ¿Cuál es el valor esperado de $\bar{p}$?
    b. ¿Cuál es el error estándar de $\bar{p}$?

c. Exprese la distribución muestral de $\bar{p}$.
d. ¿Qué indica la distribución muestral de $\bar{p}$?



32. Una proporción poblacional es 0.40. Se toma una muestra aleatoria de tamaño 200 y la proporción muestral $\bar{p}$ se usa para estimar la proporción poblacional.
    a. ¿Cuál es la probabilidad de que la proporción muestral esté entre ± 0.03 de la proporción poblacional?
    b. ¿De que la proporción muestral esté entre ± 0.05 de la proporción poblacional?
33. Suponga que la proporción poblacional es 0.55. Calcule el error estándar de la proporción, $\sigma_{\bar{p}}$, para los tamaños de muestra 100, 200, 500 y 1000. ¿Qué puede decir acerca del tamaño del error estándar a medida que el tamaño de la muestra aumenta?
34. La proporción poblacional es 0.30. ¿Cuál es la probabilidad de que las proporciones muestral y poblacional esté entre $\pm$ 0.04 con los tamaños de muestra siguientes?
    a. $n = 100$
    b. $n = 200$
    c. $n = 500$
    d. $n = 1000$
    e. ¿Qué ventaja tiene un tamaño grande de muestra?

Aplicaciones



35. El director de una empresa piensa que 30% de los pedidos provienen de nuevos compradores. Para ver la proporción de nuevos compradores se usará una muestra aleatoria simple de 100 pedidos.
    a. Suponga que el director está en lo cierto y que $p = 0.30$. ¿Cuál es la distribución muestral de $\bar{p}$ en este estudio?
    b. ¿Cuál es la probabilidad de que la proporción muestral de $\bar{p}$ esté entre 0.20 y 0.40?
    c. ¿Cuál es la probabilidad que la proporción muestral de $\bar{p}$ esté entre 0.25 y 0.35?
36. *The Cincinnati Enquirer* informa que en Estados Unidos 66% de los adultos y 87% de los jóvenes entre 12 y 17 años usan Internet (*The Cincinnati Enquirer*, 7 de febrero de 2007). Considere estos datos como proporciones poblacionales y suponga que se usará una muestra de 300 adultos y 300 jóvenes para obtener información respecto de su opinión acerca de la seguridad en Internet.
    a. Muestre la distribución muestral de $\bar{p}$, siendo $\bar{p}$ la proporción muestral de adultos que usan Internet.
    b. ¿Cuál es la probabilidad de que la diferencia entre la proporción muestral y la proporción poblacional de adultos que usan Internet no sea mayor que $\pm$ 0.04?
    c. ¿Cuál es la probabilidad de que la diferencia entre la proporción muestral y la proporción poblacional de jóvenes que usan Internet no sea mayor que $\pm$ 0.04?
    d. ¿Son diferentes las probabilidades del inciso b y del inciso c? Si es así, ¿por qué?
    e. Responda al inciso b en el caso de que el tamaño de la muestra sea 600. ¿Es menor la probabilidad? ¿Por qué?
37. Los sondeos de *Time/*CNN entre los votantes siguieron la opinión del público respecto de los candidatos presidenciales en las votaciones del 2000. En uno de estos sondeos Yankelovich Partners empleó una muestra de 589 probables votantes (*Time*, 26 de junio de 2000). Suponga que la proporción poblacional a favor de un determinado candidato a la presidencia haya sido $p = 0.50$. Sea $\bar{p}$ la proporción muestral en los posibles votantes que está a favor de ese candidato a la presidencia.
    a. Muestre la distribución muestral de $\bar{p}$.
    b. ¿Cuál es la probabilidad de que los sondeos de *Time/*CNN indiquen que la diferencia entre las proporciones muestral y poblacional en uno de estos sondeos no sea mayor que ± 0.04?
    c. ¿Cuál es la probabilidad de que los sondeos de *Time/*CNN indiquen que la diferencia entre las proporciones muestral y poblacional en uno de estos sondeos no sea mayor que ± 0.03?
    d. ¿Cuál es la probabilidad de que los sondeos de *Time/*CNN indiquen que la diferencia entre las proporciones muestral y poblacional en uno de estos sondeos no sea mayor que ± 0.02?

38. Roper ASW realizó una encuesta para obtener información acerca de la opinión de los estadounidenses respecto al dinero y la felicidad (*Money*, octubre de 2003). Cincuenta y seis por ciento de los entrevistados dijo revisar el estado de su bloc de cheques por lo menos una vez al mes.
    a. Suponga que se toma una muestra de 400 estadounidenses adultos. Indique la distribución muestral de la proporción de adultos que revisan el estado de su bloc de cheques por lo menos una vez al mes.
    b. ¿Cuál es la probabilidad de que la diferencia entre la proporción muestral y la proporción poblacional no sea mayor que ± 0.02?
    c. ¿Cuál es la probabilidad de que la diferencia entre las proporciones muestral y poblacional no sea mayor que ± 0.04?
39. El *Democrat and Chronicle* informa que 25% de los vuelos que llegaron al aeropuerto de San Diego en los primeros cinco meses de 2001, arribaron con retraso (*Democrat and Chronicle*, 23 de julio de 2001). Suponga que la proporción poblacional sea $p = 0.25$.
    a. Muestre la distribución muestral de $\bar{p}$, la proporción de vuelos retrasados en una muestra de 1 000 vuelos.
    b. ¿Cuál es la probabilidad de que la diferencia entre las proporciones muestral y poblacional no sea mayor que ± 0.03, si el tamaño de la muestra es 1000?
    c. Responda el inciso b con una muestra de 500 vuelos.
40. The Grocery Manufacturers of America informa que 76% de los consumidores leen los ingredientes que se enumeran en la etiqueta de un producto. Suponga que la proporción poblacional es $p = 0.76$ y que de la población de consumidores se selecciona una muestra de 400 consumidores.
    a. Exprese la distribución muestral de la proporción muestral $\bar{p}$, si $\bar{p}$ es la proporción de consumidores de la muestra que lee los ingredientes que se enumeran en la etiqueta.
    b. ¿Cuál es la probabilidad de que la diferencia entre las proporciones muestral y poblacional no sea mayor que ± 0.03?
    c. Conteste el inciso b si el tamaño de la muestra es 750 consumidores.
41. El Food Marketing Institute informa que 17% de los hogares gastan más de $100 en productos de abarrotes. Suponga que la proporción poblacional es $p = 0.17$ y que de la población se toma una muestra aleatoria simple de 800 hogares.
    a. Exprese la distribución muestral de $\bar{p}$, la proporción muestral de hogares que gastan más de $100 semanales en abarrotes.
    b. ¿Cuál es la probabilidad de que la proporción poblacional no difiera en más de 0.02 de la proporción poblacional?
    c. Conteste el inciso b en el caso de que el tamaño de la muestra sea 1600 hogares.

7.7 Propiedades de los estimadores puntuales

En este capítulo se ha mostrado que los estadísticos muestrales, como la media muestral $\bar{x}$, la desviación estándar muestral s y la proporción muestral $\bar{p}$ sirven como estimadores puntuales de sus correspondientes parámetros poblacionales, μ, σ y p. Resulta interesante que cada uno de estos estadísticos muestrales sean los estimadores puntuales de sus correspondientes parámetros poblacionales. Sin embargo, antes de usar un estadístico muestral como estimador puntual, se verifica si el estimador puntual tiene ciertas propiedades que corresponden a un buen estimador puntual. En esta sección se estudian las propiedades que deben tener los buenos estimadores puntuales: insesgadez, eficiencia y consistencia.

Como hay distintos estadísticos muestrales que se usan como estimadores puntuales de sus correspondientes parámetros poblacionales, en esta sección se usará la notación general siguiente.

$$\theta = \text{el parámetro poblacional de interés}$$
$$\hat{\theta} = \text{el estadístico muestral o estimador puntual de } \theta$$

En esta notación θ es la letra griega theta y la notación $\hat{\theta}$ se lee "theta sombrero". En general, θ representa cualquier parámetro poblacional como, por ejemplo, la media poblacional, la desvia-

ción estándar poblacional, la proporción poblacional, etc.; $\hat{\theta}$ representa el correspondiente estadístico muestral, por ejemplo, la media muestral, la desviación estándar muestral y la proporción muestral.

Insesgadez

Si el valor esperado del estadístico muestral es igual al parámetro poblacional que se estudia, se dice que el estadístico muestral es un *estimador insesgado* del parámetro poblacional.

INSESGADEZ

El estadístico muestral $\hat{\theta}$ es un estimado insesgado del parámetro poblacional θ si

$$E(\hat{\theta}) = \theta$$

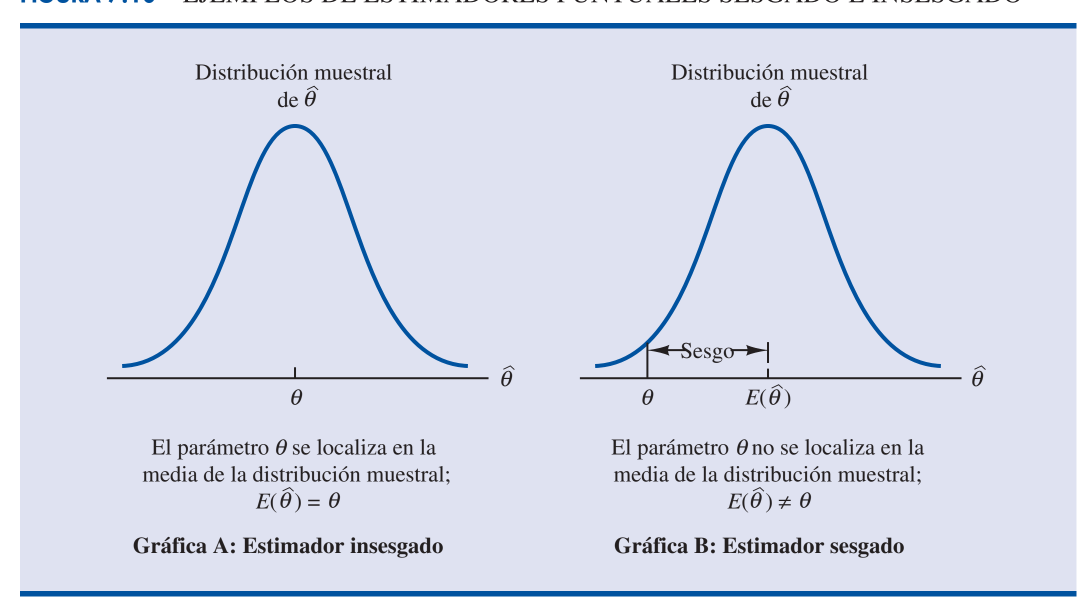
donde

$$E(\hat{\theta}) = \text{valor esperado del estadístico muestral } \hat{\theta}$$

Por tanto, el valor esperado, o media, de todos los posibles valores de un estadístico muestral insesgado es igual al parámetro poblacional que se estudia.

En la figura 7.10 se muestran los casos de los estimadores puntuales sesgado e insesgado. En la figura en que se muestra el estimador insesgado, la media de la distribución muestral es igual al valor del parámetro poblacional. En este caso los errores de estimación se equilibran, ya que algunas veces el valor del estimador puntual $\hat{\theta}$ puede ser menor que θ y otras veces sea mayor que θ. En el caso del estimador sesgado, la media de la distribución muestral es menor o mayor que el valor del parámetro poblacional. En la gráfica B de la figura 7.10, $E(\hat{\theta})$ es mayor que θ; así, la probabilidad de que los estadísticos muestrales sobreestimen el valor del parámetro poblacional es grande. En la figura se muestra la amplitud de este sesgo.

FIGURA 7.10 EJEMPLOS DE ESTIMADORES PUNTUALES SESGADO E INSESGADO



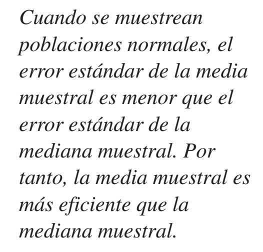
Al estudiar las distribuciones muestrales de la media muestral y de la proporción muestral, se vio que $E(\bar{x}) = \mu$ y que $E(\bar{p}) = p$. Por tanto, $\bar{x}$ y $\bar{p}$ son estimadores insesgados de sus correspondientes parámetros poblacionales μ y p.

En el caso de la desviación estándar muestral s y de la varianza muestral s^2, se puede mostrar que $E(s^2) = \sigma^2$. Por tanto, se concluye que la varianza muestral s^2 es un estimador insesgado de la varianza poblacional σ^2. En efecto, en el capítulo 3, cuando se presentaron las fórmulas para la varianza muestral y la desviación estándar muestral en el denominador se usó $n - 1$ en lugar de n para que la varianza muestral fuera un estimado insesgado de la varianza poblacional.

Eficiencia

Suponga que se usa una muestra aleatoria simple de n elementos para obtener dos estimadores puntuales insesgados de un mismo parámetro poblacional. En estas circunstancias preferirá usar el estimador puntual que tenga el menor error estándar, ya que dicho estimador tenderá a dar estimaciones más cercanas al parámetro poblacional. Se dice que el estimador puntual con menor error estándar tiene mayor **eficiencia relativa** que los otros.

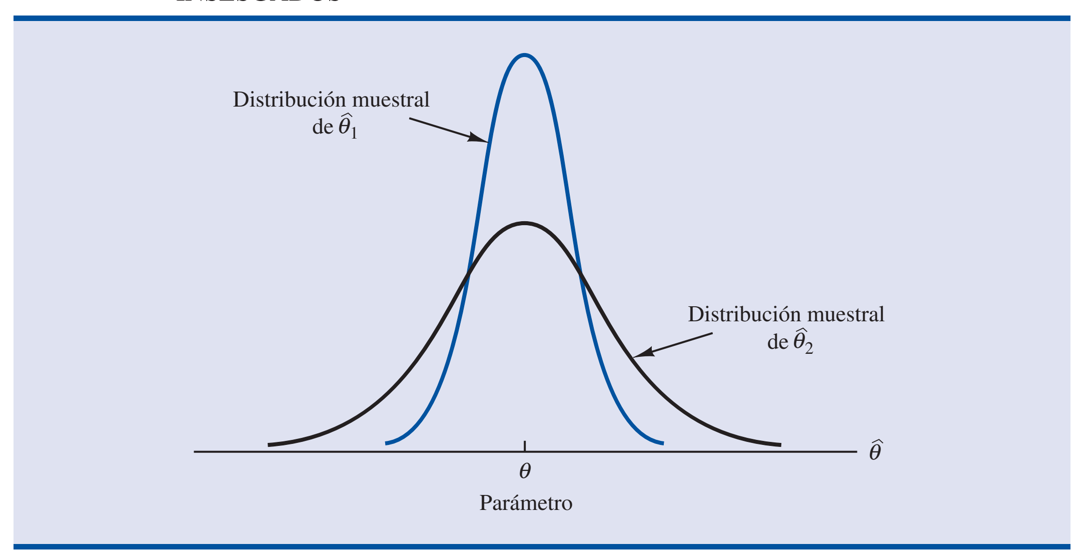
En la figura 7.11 se presentan las distribuciones muestrales de dos estimadores puntuales insesgados, $\hat{\theta}_1$ y $\hat{\theta}_2$. Observe que el error estándar de $\hat{\theta}_1$ es menor que el error estándar de $\hat{\theta}_2$; por tanto, los valores de $\hat{\theta}_1$ tienen más posibilidades de estar cerca del parámetro θ que los valores de $\hat{\theta}_2$. Como el error estándar del estimado puntual $\hat{\theta}_1$ es menor que el error estándar del estimado puntual $\hat{\theta}_2$, $\hat{\theta}_1$ es relativamente más eficiente que $\hat{\theta}_2$ y se prefiere como estimador puntual.


Cuando se muestrean poblaciones normales, el error estándar de la media muestral es menor que el error estándar de la mediana muestral. Por tanto, la media muestral es más eficiente que la mediana muestral.

Consistencia

La tercera propiedad relacionada con un buen estimador puntual es la **consistencia**. Dicho de manera sencilla, un estimador puntual es consistente si el valor del estimador puntual tiende a estar más cerca del parámetro poblacional a medida que el tamaño de la muestra aumenta. En otras palabras, una muestra grande tiende a proporcionar mejor estimación puntual que una pequeña. Observe que en el caso de la media muestral $\bar{x}$, el error estándar de $\bar{x}$ está dado por $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Puesto que $\sigma_{\bar{x}}$ está vinculado con el tamaño de la muestra, de manera que muestras mayores dan

FIGURA 7.11 DISTRIBUCIONES MUESTRALES DE DOS ESTIMADORES PUNTUALES INSESGADOS



valores menores de $\sigma_{\bar{x}}$, entonces muestras de tamaño grande tienden a proporcionar estimadores puntuales más cercanos a la media poblacional μ. Mediante un razonamiento similar, concluya que la proporción muestral $\bar{p}$ es un estimador consistente de la proporción poblacional p.

NOTAS Y COMENTARIOS

En el capítulo 3 se dijo que la media y la mediana son dos medidas de localización central. En este capítulo sólo se estudió la media. La razón es que cuando se muestrea de una población normal, en la cual la media y la mediana poblacionales son idénticas, el error estándar de la mediana es cerca de 25% mayor que el error estándar de la media. Recuerde que en el problema de EAI con $n = 30$, el error estándar de la media fue $\sigma_{\bar{x}} = 730.3$. El error estándar de la mediana en este problema será $1.25 \times (730.7) = 913$. Por tanto, la media muestral es más eficiente y tendrá más probabilidad de estar dentro de una determinada distancia de la media poblacional.



Otros métodos de muestreo

Esta sección proporciona una breve introducción a otros métodos de muestreo distintos al muestreo aleatorio simple.

Se describió el procedimiento de muestreo aleatorio simple y se estudiaron las propiedades de las distribuciones muestrales de $\bar{x}$ y de $\bar{p}$ cuando se usa el muestreo aleatorio simple. Sin embargo, el muestreo aleatorio simple no es el único método de muestreo que existe. Hay otros métodos como el muestro aleatorio estratificado, el muestreo por conglomerados y el muestreo sistemático que, en ciertas situaciones, tienen ventajas sobre el muestreo aleatorio simple. En esta sección se introducen brevemente estos métodos de muestreo. En el capítulo 22 que se encuentra en el CD que se distribuye con el texto se estudian estos métodos de muestreo con más detenimiento.

Muestreo aleatorio estratificado

El muestreo aleatorio estratificado funciona mejor cuando la varianza entre los elementos de cada estrato es relativamente pequeña.

En el **muestreo aleatorio estratificado** los elementos de la población primero se dividen en grupos, a los que se les llama *estratos*, de manera que cada elemento pertenezca a uno y sólo un estrato. La base para la formación de los estratos, que puede ser departamento, edad, tipo de industria, etc., está a discreción de la persona que diseña la muestra. Sin embargo, se obtienen mejores resultados cuando los elementos que forman un estrato son lo más parecido posible. La figura 7.12 es un diagrama de una población dividida en H estratos.

Una vez formados los estratos, se toma una muestra aleatoria simple de cada estrato. Existen fórmulas para combinar los resultados de las muestras de los varios estratos en una estimación

FIGURA 7.12 DIAGRAMA DE UN MUESTREO ALEATORIO ESTRATIFICADO

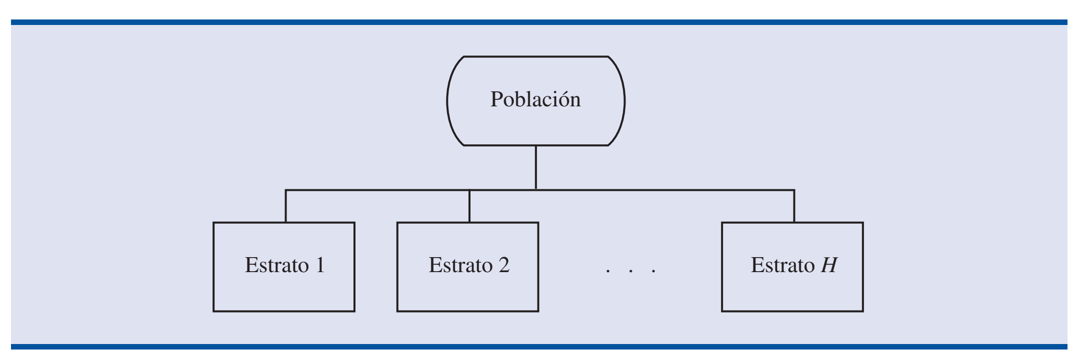
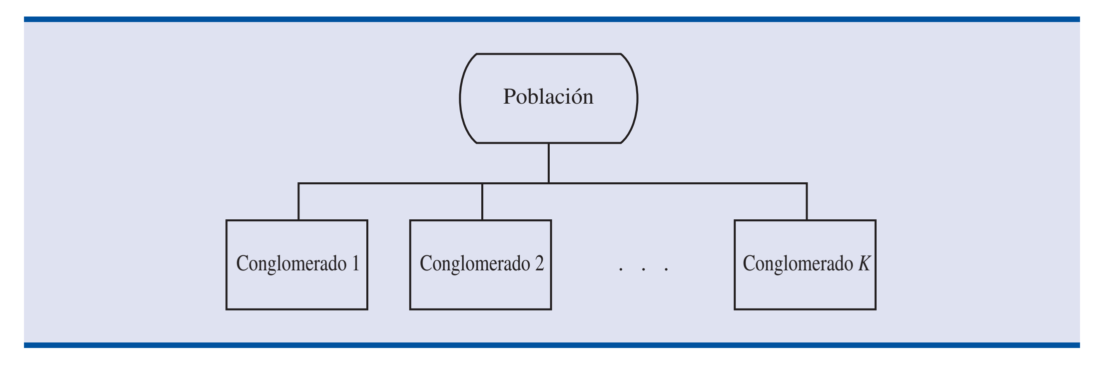


FIGURA 7.13 DIAGRAMA DEL MUESTREO POR CONGLOMERADOS



del parámetro poblacional de interés. El valor del muestreo aleatorio estratificado depende de qué tan homogéneos sean los elementos dentro de cada estrato. Si los elementos de un estrato son homogéneos, el estrato tendrá una varianza pequeña. Por tanto, con muestras relativamente pequeñas de los estratos se obtienen buenas estimaciones de las características de los estratos. Si los estratos son homogéneos, el muestreo aleatorio estratificado, proporciona resultados tan precisos como los de un muestreo aleatorio simple, pero con una muestra de tamaño total menor.

Muestreo por conglomerados

El muestreo por conglomerados funciona mejor cuando cada conglomerado proporciona una representación a menor escala de la población.

En el **muestreo por conglomerados** los elementos de la muestra primero se dividen en grupos separados, llamados *conglomerados.* Cada elemento de la población pertenece a uno y sólo un conglomerado (véase figura 7.13). Se toma una muestra aleatoria simple de los conglomerados. La muestra está formada por todos los elementos dentro de cada uno de los conglomerados que forman la muestra. El muestreo por conglomerados tiende a proporcionar mejores resultados cuando los elementos dentro de los conglomerados no son semejantes. Lo ideal es que cada conglomerado sea una representación, a pequeña escala, de la población. Si todos los conglomerados son semejantes en este aspecto, tomando en la muestra un número pequeño de conglomerados se obtendrá una buena estimación de los parámetros poblacionales.

Una de las principales aplicaciones del muestro por conglomerados es el muestreo de áreas, en el que los conglomerados son las manzanas de una ciudad u otras áreas bien definidas. El muestreo por conglomerados requiere, por lo general, tamaños de muestra mayores que los requeridos en el muestreo aleatorio simple o en el muestreo aleatorio estratificado. Sin embargo, es posible reducir costos debido a que cuando se envía a un entrevistador a uno de los conglomerados de la muestra (por ejemplo, a una manzana de una ciudad), es posible obtener muchas observaciones en poco tiempo. Por tanto, se obtiene una muestra de tamaño grande a un costo significantemente menor.

Muestreo sistemático

Para ciertos muestreos, en especial en aquellos con poblaciones grandes, se necesita mucho tiempo para tomar una muestra aleatoria simple (hallando primero los números aleatorios y después contando y recorriendo toda una lista de la población hasta encontrar los elementos correspondientes). Una alternativa al muestreo aleatorio simple es el **muestreo sistemático**. Por ejemplo, si se quiere una muestra de tamaño 50 de una población que tiene 5000 elementos, se muestrea uno de cada 5 000/50 = 100 elementos de la población. En este caso, un muestreo sistemático consiste en seleccionar en forma aleatoria uno de los primeros elementos de la lista de la población. Los otros elementos se identifican contando a partir del primer elemento 100 elementos para tomar el elemento que tenga la posición 100 en la lista de la población, a partir de este elemento se cuentan otros 100 y así se continúa. Por lo general, de esta manera es más fácil de identificar la muestra de 50 que si se usara el muestreo aleatorio simple. Como el primer elemento que se selecciona es elegido en forma aleatoria, se supone que una muestra sistemática tiene las

propiedades de una muestra aleatoria simple. Esta suposición es aplicable, en especial, cuando la lista de los elementos de la población es un orden aleatorio de los elementos.

Muestreo de conveniencia

Los métodos de muestreo hasta ahora vistos se conocen como técnicas *probabilísticas de muestreo.* Los elementos seleccionados de una población tienen una probabilidad conocida de ser incluidos en la muestra. La ventaja del muestreo probabilístico es que, por lo general, se identifica la distribución muestral del estadístico muestral correspondiente. Para determinar las propiedades de la distribución muestral se usan las fórmulas presentadas en este capítulo para el muestreo aleatorio simple. La distribución muestral permite hacer afirmaciones probabilísticas acerca del error al usar los resultados muestrales para hacer inferencias acerca de la población.

El **muestreo de conveniencia** es una técnica de *muestreo no probabilística*. Como el nombre lo indica, la muestra se determina por conveniencia. Los elementos se incluyen en la muestra sin que haya una probabilidad previamente especificada o conocida de que sean incluidos en la muestra. Por ejemplo, un profesor que realiza una investigación en una universidad puede usar estudiantes voluntarios para que constituyan una muestra; ¿la razón para elegirlos? simple, los tiene al alcance y participarán como sujetos a un costo bajo o sin costo. De manera similar, un inspector puede muestrear un cargamento de naranjas seleccionando al azar naranjas de varias de las cajas. Marcar cada naranja y usar un método probabilístico de muestreo puede no resultar práctico. Muestras como los paneles de voluntarios en investigaciones sobre los consumidores son también muestras de conveniencia.

Dichas muestras tienen la ventaja de que es relativamente fácil seleccionar la muestra y recoger los datos; sin embargo, es imposible evaluar la “bondad” de la muestra en términos de su representatividad de la población. Una muestra de conveniencia puede o no dar buenos resultados. Algunas veces los investigadores aplican los métodos estadísticos propios de muestras probabilísticas a las muestras de conveniencia, con el argumento de que la muestra de conveniencia se trata como si fuera una muestra probabilística. Sin embargo, estos argumentos no tienen fundamento y se debe tener cuidado al interpretar los resultados de muestreos de conveniencia que han sido usados para hacer inferencias acerca de la población.

Muestreo subjetivo

Otra técnica de muestreo no probabilística es el muestreo subjetivo. En este método la persona que más sabe sobre un asunto selecciona elementos de la población que considera los más representativos de la población. Este método suele ser una manera relativamente fácil de seleccionar una muestra. Por ejemplo, un reportero puede seleccionar dos o tres senadores considerando que estos senadores reflejan la opinión general de todos los senadores. Sin embargo, la calidad de los resultados muestrales depende de la persona que selecciona la muestra. Aquí también hay que tener mucho cuidado al hacer inferencias acerca de las poblaciones a partir de muestreos subjetivos.

NOTAS Y COMENTARIOS

Se recomienda usar métodos de muestreo probabilístico: muestro aleatorio simple, muestreo aleatorio estratificado, muestreo por conglomerados o muestreo sistemático. Si se usan estos métodos existen fórmulas para evaluar la “bondad” de los resultados muestrales en términos de la cercanía de los resultados a los parámetros poblacionales que se estiman. Con los muestreos de conveniencia o con los muestreos subjetivos no se puede estimar la bondad de los resultados. Por tanto, debe tenerse mucho cuidado al interpretar resultados basados en métodos de muestreo no probabilístico.

Resumen

En este capítulo se presentaron los conceptos de muestreo aleatorio simple y de distribución muestral. Se mostró cómo seleccionar una muestra aleatoria simple y la forma de usar los datos recolectados de la muestra para obtener estimadores puntuales de los parámetros poblacionales. Ya que distintas muestras aleatorias simples dan valores diferentes de los estimadores puntuales, los estimadores puntuales como $\bar{x}$ y $\bar{p}$ son variables aleatorias. A la distribución de probabilidad de una variable aleatoria de este tipo se le conoce como distribución muestral. En particular, se describieron la distribución muestral de la media muestral $\bar{x}$ y la distribución muestral de la proporción muestral $\bar{p}$.

Al estudiar las características de las distribuciones muestrales de $\bar{x}$ y de $\bar{p}$, se vio que $E(\bar{x}) = \mu$ y que $E(\bar{p}) = p$. Después de dar las fórmulas para la desviación estándar o error estándar de dichos estimadores, se describieron las condiciones necesarias para que las distribuciones muestrales de $\bar{x}$ y de $\bar{p}$ sigan una distribución normal. Otros métodos de muestreo fueron el muestreo aleatorio estratificado, el muestreo por conglomerados, el muestreo sistemático, el muestreo por conveniencia y el muestreo subjetivo.

Glosario

Parámetro Característica numérica de una población, por ejemplo, la media poblacional μ, la desviación estándar poblacional σ, la proporción poblacional p, etcétera.

Muestreo aleatorio simple Poblaciones finitas: muestra seleccionada de manera que cada una de las muestras de tamaño n tenga la misma probabilidad de ser seleccionada. Poblaciones infinitas: muestra seleccionada de manera que todos los elementos provengan de la misma población y los elementos se seleccionen de manera independiente.

Muestreo sin reemplazo Una vez que un elemento ha sido incluido en la muestra, se retira de la población y ya no se selecciona una vez más.

Muestreo con reemplazo Una vez que un elemento se ha incluido en la muestra, se regresa a la población. Un elemento ya seleccionado para la muestra puede ser seleccionado nuevamente y puede aparecer más de una vez en la muestra.

Estadístico muestral Característica muestral, por ejemplo, la media muestral $\bar{x}$, la desviación estándar muestral s, la proporción muestral $\bar{p}$, etc. El valor del estadístico muestral se usa para estimar el valor del correspondiente parámetro poblacional.

Estimador puntual Un estadístico muestral como $\bar{x}$, s, o $\bar{p}$ que proporciona una estimación puntual del parámetro poblacional correspondiente.

Estimación puntual Valor de un estimador que se usa en una situación particular como estimación del parámetro poblacional.

Distribución muestral Distribución de probabilidad que consta de todos los posibles valores de un estadístico muestral.

Insesgado Propiedad de un estimador que consiste en que el valor esperado del estimador puntual es igual al parámetro poblacional que estima.

Factor de corrección para una población finita Es el factor $\sqrt{(N-n)/(N-1)}$ que se usa en las fórmulas de $\sigma_{\bar{x}}$ y $\sigma_{\bar{p}}$ siempre que se muestrea de una población finita y no de una población infinita. Sin embargo, hay una regla generalmente aceptada, ignorar el factor de corrección en una población finita siempre que $n/N \leq 0.05$.

Error estándar La desviación estándar de un estimador puntual.

Teorema del límite central Permite usar la distribución de probabilidad normal para aproximar la distribución muestral de $\bar{x}$ siempre que la muestra sea grande.

Eficiencia relativa Dados dos estimadores puntuales insesgados de un mismo parámetro poblacional, el estimador puntual que tenga menor error estándar será más eficiente.

Consistencia Propiedad de un estimador puntual que está presente siempre que muestras más grandes dan estimaciones puntuales más cercanas al parámetro poblacional.
Muestreo aleatorio estratificado Método probabilístico en el que primero se divide la población en estratos y después se toma una muestra aleatoria simple de cada estrato.
Muestreo por conglomerados Método probabilístico en el que primero se divide la población en conglomerados y después se toma una muestra aleatoria de los conglomerados.
Muestreo sistemático Método probabilístico en el que primero se selecciona uno de los primeros k elementos de una población y después se selecciona cada k-ésimo elemento de la población.
Muestreo de conveniencia Método no-probabilístico en el que la selección de los elementos para la muestra es de acuerdo con la conveniencia.
Muestreo subjetivo Método no-probabilístico en el que la selección de los elementos para la muestra es de acuerdo con la opinión de la persona que hace el estudio.

Fórmulas clave

Valor esperado de $\bar{x}$

$$E(\bar{x}) = \mu \tag{7.1}$$

Desviación estándar de $\bar{x}$ (error estándar)

Población finita — *Población infinita*

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{7.2}$$

Valor esperado de $\bar{p}$

$$E(\bar{p}) = p \tag{7.4}$$

Desviación estándar de $\bar{p}$ (error estándar)

Población finita — *Población infinita*

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{7.5}$$

Ejercicios complementarios

42. *BusinessWeek's* Corporate Scoreboard proporciona datos trimestrales sobre distintos aspectos de las acciones de 899 empresas (*BusinessWeek*, 14 de agosto de 2000). Las empresas son numeradas del 1 al 899 en el orden en que aparecen en la lista del Corporate Scoreboard. Remítase a la parte inferior de la segunda columna de dígitos aleatorios de la tabla 7.1, ignore los dos primeros dígitos de cada conjunto de números, use números de tres dígitos, empiece con el 112, lea hacia arriba de la columna para determinar las ocho primeras empresas a incluir en una muestra aleatoria simple.

43. Los estadounidenses están cada vez más preocupados por el aumento en los costos de Medicare. En 1990 el promedio de gastos anuales de un derechohabiente de Medicare era \$3267; en el 2003 el promedio de gastos anuales de un derechohabiente de Medicare era \$6883 (*Money*, otoño de

2003). Suponga que usted contrata a una empresa consultora para tomar una muestra de 50 de los derechohabientes de Medicare en 2003 con objeto de investigar los gastos. Asuma que la desviación estándar en los gastos de Medicare en 2003 haya sido de $2000.

a. Muestre la distribución muestral de la media, en muestras de tamaño cincuenta, de los gastos de derechohabientes de Medicare en 2003.
b. ¿Cuál es la probabilidad de que la media muestral no se aleje más de $\pm$$300 de la media poblacional?
c. ¿Cuál es la probabilidad de que la media muestral sea mayor que $7500? Si la empresa que contrató le dice que la media muestral en los derechohabientes que entrevistó es $7500, ¿dudaría que la empresa contratada hubiera hecho un muestreo adecuado? ¿Por qué sí o por qué no?

44. *BusinessWeek* encuesta a ex alumnos de administración 10 años después de terminados sus estudios (*BusinessWeek*, 22 de septiembre de 2003). Uno de los hallazgos fue que gastan en promedio $115.50 semanales en comidas sociales. A usted se le pide que realice un estudio con una muestra de 40 de estos ex alumnos.
    a. Muestre la distribución muestral de $\bar{x}$, la media muestral de los gastos de 40 ex alumnos.
    b. ¿Cuál es la probabilidad de que la media muestral no se aleje en más o menos de $10 de la media poblacional?
    c. Suponga que encuentra una media muestral de $100. ¿Cuál es la probabilidad de hallar una media muestral de $100 o menos? ¿Consideraría que los ex alumnos de esta muestra son un grupo inusual respecto a estos gastos? ¿Por qué sí o por qué no?

45. El tiempo promedio que un estadounidense ve televisión es 15 horas por semana (*Money*, noviembre de 2003). Suponga que se toma una muestra de 60 estadounidenses para investigar con más detalle sus hábitos a este respecto. Asuma que la desviación estándar poblacional en las horas de televisión semanales es $\sigma = 4$ horas.
    a. ¿Cuál es la probabilidad de que la media muestral no se aleje más o menos de 1 hora de la media poblacional?
    b. ¿Cuál es la probabilidad de que la media muestral no se aleje más o menos de 45 minutos de la media poblacional?

46. En Indiana el salario anual promedio de un empleado del gobierno federal es $41 979 (*The World Almanac*, 2001). Use esta cifra como media poblacional y suponga que la desviación estándar poblacional es $\sigma = \$5000$. Suponga que se selecciona una muestra de 50 de estos empleados del gobierno federal.
    a. ¿Cuál es el valor del error estándar de la media?
    b. ¿Cuál es la probabilidad de que la media muestral sea mayor que $41 979?
    c. ¿Cuál es la probabilidad de que la media muestral no se aleje más o menos de $1000 de la media poblacional?
    d. ¿Qué tanto variaría la probabilidad del inciso c si el tamaño de la muestra se aumentara a 100?

47. Tres empresas llevan inventarios de distintos tamaños. El inventario de la empresa A contiene 2000 artículos, el inventario de la empresa B tiene 5000 artículos y el inventario de la empresa C 10 000. La desviación estándar poblacional de los costos de los artículos en los inventarios de estas empresas es $\sigma = 144$. Un consultor de estadística recomienda que cada empresa tome una muestra de 50 artículos de su inventario para obtener una estimación estadística válida del costo promedio por artículo. Los administradores de la empresa más pequeña opinan que como su población es menor se podrá hacer la estimación con una muestra mucho más pequeña de la que se requiere para la empresa más grande. Sin embargo, el consultor opina que para tener el mismo error estándar y, por tanto, la misma precisión en los resultados muestrales, todas las empresas deberán emplear el mismo tamaño de muestra, sin importar el tamaño de la población.
    a. Con el factor de corrección para una población finita, calcule el error estándar de cada una de las tres empresas para un tamaño de muestra de 50.
    b. ¿Cuál es la probabilidad en cada empresa de que la media muestral $\bar{x}$ esté a no más de ±25 de la media poblacional μ?

48. Un investigador informa sobre sus resultados diciendo que el error estándar de la media es 20. La desviación estándar poblacional es 500.
    a. ¿De qué tamaño fue la muestra usada en esta investigación?
    b. ¿Cuál es la probabilidad de que la estimación puntual esté a no más de ±25 de la media poblacional?

49. Un inspector de control de calidad vigila periódicamente un proceso de producción. El inspector selecciona muestras aleatorias simples de artículos ya terminados y calcula la media muestral del peso del producto $\bar{x}$. Si en un periodo largo se encuentra que 5% de los valores de $\bar{x}$ son mayores que 2.1 libras y 5% son menores que 1.9 libras. ¿Cuáles son la media y la desviación estándar de la población de los productos elaborados en este proceso?

50. Cerca de 28% de las empresas tienen como propietario a una mujer (*The Cincinnati Enquirer*, 26 de enero de 2006). Responda estas preguntas con base en una muestra de 240 empresas.
    a. Muestre la distribución muestral de $\bar{p}$, la proporción muestral de las empresas propiedad de una mujer.
    b. ¿Cuál es la probabilidad de que la proporción muestral esté a no más de ±0.04 de la proporción poblacional?
    c. ¿Cuál es la probabilidad de que la proporción muestral esté a no más de ±0.02 de la proporción poblacional?

51. Una empresa de investigación de mercado realiza encuestas telefónicas con una tasa de respuesta de 40%, de acuerdo con la experiencia. ¿Cuál es la probabilidad de que en una muestra de 400 números telefónicos 150 personas cooperen y respondan las preguntas? En otras palabras, ¿cuál es la probabilidad de que la proporción muestral sea al menos $150/400 = 0.375$?

52. Los publicistas contratan proveedores de servicios de Internet y motores de búsqueda para poner su publicidad en los sitios Web. Pagan una cuota de acuerdo con el número de clientes potenciales que hacen clic en su publicidad. Por desgracia, el fraude por clic —la práctica de hacer clic en una publicidad con el solo objeto de aumentar las ganancias— se ha convertido en un problema. Cuarenta por ciento de los publicistas se quejan de haber sido víctima de fraude por clic (*BusinessWeek*, 13 de marzo de 2006). Suponga que se toma una muestra aleatoria de 380 publicistas con objeto de tener más información acerca de cómo son afectados por este fraude por clic.
    a. ¿Cuál es la probabilidad de que la proporción muestral esté a no más de ±0.04 de la proporción poblacional?
    b. ¿Cuál es la probabilidad de que la proporción muestral sea mayor que 0.45?

53. La proporción de personas aseguradas con una compañía de seguros para automóviles que tienen una multa de tráfico en el periodo de un año es 0.15
    a. Indique la distribución muestral de $\bar{p}$ si se emplea una muestra aleatoria de 150 asegurados para determinar la proporción de quienes han tenido por lo menos una multa en un año.
    b. ¿Cuál es la probabilidad de que la proporción muestral esté a no más de ±0.03 de la proporción poblacional?

54. Lori Jeffrey es un exitoso representante de ventas de libros universitarios, tiene éxito en sus recomendaciones de libros en 25% de sus llamadas. Considere sus llamadas de ventas de un mes como muestra de todas sus posibles llamadas, suponga que en el análisis estadístico de los datos se encuentra que el error estándar de la proporción es 0.0625.
    a. ¿De qué tamaño fue la muestra que se usó en el análisis? Es decir, ¿cuántas llamadas hizo Lori Jeffrey en ese mes?
    b. Sea $\bar{p}$ la proporción muestral de éxitos en sus recomendaciones de libros en ese mes. Muestre la distribución muestral de $\bar{p}$.
    c. Mediante la distribución muestral de $\bar{p}$, calcule la probabilidad de que el vendedor tenga éxito en 30% o más de las llamadas de ventas en el lapso de un mes.

Apéndice 7.1 Valor esperado y desviación estándar de $\bar{x}$

En este apéndice se presentan las bases matemáticas de las expresiones $E(\bar{x})$, valor esperado de $\bar{x}$, ecuación (7.1), y $\sigma_{\bar{x}}$, desviación estándar de $\bar{x}$, ecuación (7.2).

Valor esperado de $\bar{x}$

Se tiene una población que tiene media μ y varianza σ^2. Se selecciona una muestra aleatoria de tamaño n cuyas observaciones se denotan $x_1, x_2, \ldots, x_n$. La media muestral $\bar{x}$ se calcula como sigue.

$$\bar{x} = \frac{\Sigma x_i}{n}$$

Si se repiten los muestreos aleatorios de tamaño n, $\bar{x}$ será una variable aleatoria que tomará diferentes valores dependiendo de los n elementos que formen la muestra. El valor esperado de la variable aleatoria $\bar{x}$ es la media de todos los posibles valores $\bar{x}$.

$$\begin{aligned}
\text{Media de } \bar{x} = E(\bar{x}) &= E\left(\frac{\Sigma x_i}{n}\right) \\
&= \frac{1}{n}[E(x_1 + x_2 + \cdots + x_n)] \\
&= \frac{1}{n}[E(x_1) + E(x_2) + \cdots + E(x_n)]
\end{aligned}$$

Para cada x_i se tiene $E(x_i) = \mu$; por tanto,

$$\begin{aligned}
E(\bar{x}) &= \frac{1}{n}(\mu + \mu + \cdots + \mu) \\
&= \frac{1}{n}(n\mu) = \mu
\end{aligned}$$

Este resultado indica que la media de todos los posibles valores de $\bar{x}$ es igual a la media poblacional μ. Es decir $E(\bar{x}) = \mu$.

Desviación estándar de $\bar{x}$

Se tiene, de nuevo, una población con media μ y varianza σ^2 y una media muestral dada por

$$\bar{x} = \frac{\Sigma x_i}{n}$$

Se sabe que $\bar{x}$ es una variable aleatoria que toma distintos valores en distintas muestras aleatorias de tamaño n, dependiendo de los elementos que constituyen la muestra. Lo que sigue es una deducción de la fórmula para la desviación estándar de los valores de $\bar{x}$, $\sigma_{\bar{x}}$, en el caso en el que la población sea infinita. La deducción de la fórmula para $\sigma_{\bar{x}}$ cuando la población es finita y el muestreo se hace sin reemplazo es más complicada y queda fuera de los alcances de este texto.

De regreso al caso de una población infinita, recuerde que una muestra aleatoria simple de una población infinita, consta de observaciones $x_1, x_2, \ldots, x_n$ que son independientes. Las dos expresiones siguientes son fórmulas generales para la varianza de una variable aleatoria.

$$\text{Var}(ax) = a^2 \, \text{Var}(x)$$

donde a es una constante y x es una variable aleatoria, y

$$\text{Var}(x + y) = \text{Var}(x) + \text{Var}(y)$$

donde x y y son variables aleatorias *independientes*. Usando las ecuaciones anteriores, se puede deducir la fórmula para la varianza de la variable $\bar{x}$ como sigue.

$$\text{Var}(\bar{x}) = \text{Var}\left(\frac{\Sigma x_i}{n}\right) = \text{Var}\left(\frac{1}{n}\Sigma x_i\right)$$

Entonces, como $1/n$ es una constante, se tiene

$$\begin{aligned} \text{Var}(\bar{x}) &= \left(\frac{1}{n}\right)^2 \text{Var}(\Sigma x_i) \\ &= \left(\frac{1}{n}\right)^2 \text{Var}(x_1 + x_2 + \cdots + x_n) \end{aligned}$$

En el caso de una población infinita, las variables aleatorias $x_1, x_2, \ldots, x_n$ son independientes, lo que nos permite escribir

$$\text{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 [\text{Var}(x_1) + \text{Var}(x_2) + \cdots + \text{Var}(x_n)]$$

Para toda x_i, se tiene $\text{Var}(x_i) = \sigma^2$; por tanto se tiene

$$\text{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \cdots + \sigma^2)$$

Como en esta expresión hay n valores σ^2, se tiene

$$\text{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 (n\sigma^2) = \frac{\sigma^2}{n}$$

Sacando ahora la raíz cuadrada, se obtiene la fórmula de la desviación estándar de $\bar{x}$.

$$\sigma_{\bar{x}} = \sqrt{\text{Var}(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$

Apéndice 7.2 Muestreo aleatorio con Minitab

Si en un archivo se encuentra una lista con los elementos de una población, se puede usar Minitab para seleccionar una muestra aleatoria simple. Por ejemplo, en la columna 1 del conjunto de datos MetAreas se proporciona una lista de las 100 principales áreas metropolitanas de Estados Unidos y Canadá (*Places Rated Almanac-The Millenium Edition 2000*). La columna 2 contiene

TABLA 7.6 PUNTUACIÓN GENERAL PARA LAS PRIMERAS 10 ÁREAS METROPOLITANAS EN EL CONJUNTO DE DATOS METAREAS



Área metropolitana	Puntuación
Albany, NY	64.18
Albuquerque, NM	66.16
Appleton, WI	60.56
Atlanta, GA	69.97
Austin, TX	71.48
Baltimore, MD	69.75
Birmingham, AL	69.59
Boise City, ID	68.36
Boston, MA	68.99
Buffalo, NY	66.10

la puntuación general dada a cada área. En la tabla 7.6 se presentan las primeras 10 áreas metropolitanas con sus puntuaciones correspondientes.

Suponga que pretende seleccionar una muestra aleatoria simple de 30 áreas metropolitanas con objeto de hacer un estudio sobre el costo de la vida en Estados Unidos y Canadá. Para seleccionar la muestra aleatoria se siguen los pasos que se indica a continuación.

Paso 1. Seleccionar el menú desplegable **Calc**
Paso 2. Elegir **Random Data**
Paso 3. Elegir **Sample From Columns**
Paso 4. Cuando aparezca el cuadro de diálogo **Sample From Columns**:
Ingresar 30 en el cuadro **Sample**
Ingresar C1 C2 en el cuadro que se encuentra debajo
Ingresar C3 C4 en el cuadro **Store samples in**
Paso 5. Hacer clic en **OK**

La muestra aleatoria con las 30 áreas metropolitanas aparece en las columnas C3 y C4.

Apéndice 7.3 Muestreo aleatorio con Excel

Si en un archivo se encuentra una lista con los elementos de una población, Excel se podrá usar para seleccionar una muestra aleatoria simple. Por ejemplo, en la columna A del conjunto de datos MetAreas se proporciona una lista de las 100 principales áreas metropolitanas de Estados Unidos y Canadá (*Places Rated Almanac-The Millenium Edition 2000*).

La columna B contiene la puntuación general dada a cada área. En la tabla 7.6 se presentan las primeras 10 áreas metropolitanas con sus puntuaciones correspondientes.

Suponga que quiere seleccionar una muestra aleatoria simple de 30 áreas metropolitanas con objeto de hacer un estudio sobre el costo de la vida en Estados Unidos y Canadá.

Los renglones de cualquier conjunto de datos en Excel se pueden colocar en orden aleatorio agregando una columna al conjunto de datos y llenando la columna con números aleatorios mediante la función = ALEATORIO(); después con la herramienta de Excel Orden ascendente aplicada a la columna de números aleatorios, los renglones del conjunto de datos quedarán reordenados aleatoriamente. La muestra aleatoria de tamaño n aparecerá en los n primeros renglones del conjunto de datos reordenado.

En el conjunto de datos MetAreas, los encabezados aparecen en el renglón 1 y las 100 áreas metropolitanas se encuentran en los renglones 2 a 101. Para seleccionar una muestra aleatoria de 30 áreas metropolitanas siga los pasos siguientes.

Paso 1. Ingresar = ALEATORIO() en la celda C2.
Paso 2. Copiar la celda C2 a las celdas C3:C101
Paso 3. Seleccionar cualquier celda de la columna C
Paso 4. Clic en el botón **Orden ascendente** de la barra de herramientas.

La muestra aleatoria con 30 áreas metropolitanas aparecerá en los renglones 2 a 31 del conjunto de datos reordenado. Los números aleatorios de la columna C ya no se necesitan y pueden borrarse si se desea.

CAPÍTULO 8



Estimación por intervalo

CONTENIDO



LA ESTADÍSTICA *en* LA PRÁCTICA

FOOD LION*
SALISBURY, CAROLINA DEL NORTE

Fundada en 1957 como Food Town, Food Lion es una de las cadenas de supermercados más grande de Estados Unidos, con 1 200 tiendas en 11 estados. La empresa vende más de 24 000 productos diferentes y ofrece mercancías de marcas nacionales y regionales, así como una cantidad cada vez mayor de productos de gran calidad de marca propia fabricados especialmente. La empresa mantiene su liderazgo en precios bajos y asegura la calidad a partir de eficientes controles como formatos estándar de tienda, diseño innovador de los almacenes, instalaciones eficaces y sincronización de datos con los proveedores. Food Lion mira hacia un futuro de continua innovación, crecimiento, liderazgo de precio y servicio a sus clientes.

Puesto que es un negocio con inventario intensivo, Food Lion decidió adoptar como método de evaluación de inventario el método LIFO (*last-in*, *first-out*). Este método compara los costos corrientes con los ingresos corrientes. Lo cual minimiza los efectos de los cambios radicales de precios sobre los resultados de ganancia y pérdida. Además, el método LIFO reduce la ganancia neta disminuyendo con esto los impuestos sobre la renta durante los periodos de inflación.

Food Lion establece un índice LIFO para cada uno de los siete grupos de inventario: abarrotes, papel/artículos para el hogar, artículos para mascotas, artículos para la salud y la belleza, lácteos, cigarros/tabaco y cervezas/vinos. Por ejemplo, un índice LIFO de 1.008 para el grupo abarrotes indica que el valor de este inventario, a los costos corrientes, refleja un aumento de 0.8% debido a la inflación en el último periodo de un año.

Un índice LIFO para cada grupo de inventario requiere que el inventario de fin de año de cada producto se evalúe al costo actual de fin de año y al costo del año anterior.

*Los autores agradecen a Keith Cunningham director de impuestos y a Bobby Harkey del equipo de contadores por proporcionar este artículo para *La estadística en la práctica.*



El almacén Food Lion en el centro comercial Cambridge, Charlotte, North Carolina. © Cortesia de Food Lion.

Para ahorrar tiempo y gastos excesivos por el conteo para el inventario en las 1200 tiendas, Food Lion selecciona una muestra aleatoria simple de 50 tiendas. El inventario físico de fin de año se realiza en cada una de las tiendas de la muestra. Para obtener el índice LIFO de cada uno de los grupos de inventario se usan los costos del año actual y los costos del año anterior.

En uno de los últimos años, la estimación muestral del índice LIFO para el inventario del grupo de productos para la salud y la belleza fue 1.015. Con un nivel de confianza de 95%, Food Lion calculó un margen de error de 0.006 para la estimación muestral. Por tanto, el intervalo de 1.009 a 1.021 proporciona una estimación por intervalo de confianza de 95% del índice LIFO poblacional. Este nivel de precisión se consideró muy bueno.

En ese capítulo aprenderá cómo calcular un margen de error para una estimación puntual. También verá cómo usar esta información para construir e interpretar estimaciones por intervalo para una media poblacional y para una proporción poblacional.

En el capítulo 7 se dijo que un estimador puntual es un estadístico muestral que se usa para estimar un parámetro poblacional. Por ejemplo, la media muestral $\bar{x}$ es un estimador puntual de la media poblacional μ y la proporción muestral $\bar{p}$ es un estimador puntual de la proporción poblacional p. Como no se puede esperar que un estimador puntual suministre el valor exacto del parámetro poblacional, se suele calcular una **estimación por intervalo** al sumar y restar al estimador puntual una cantidad llamada **margen de error**. La fórmula general de una estimación por intervalo es

$$\text{Estimación puntual} \pm \text{Margen de error}$$

El objetivo de la estimación por intervalo es aportar información de qué tan cerca se encuentra la estimación puntual, obtenida de la muestra, del valor del parámetro poblacional.

En este capítulo se muestra cómo obtener una estimación por intervalo para la media poblacional μ y para la proporción poblacional p. La fórmula general para obtener una estimación por intervalo para la media poblacional es

$$\bar{x} \pm \text{Margen de error}$$

De manera similar, la fórmula general para obtener una estimación por intervalo para la proporción poblacional es

$$\bar{p} \pm \text{Margen de error}$$

Las distribuciones muestrales de $\bar{x}$ y de $\bar{p}$ son clave para calcular estas estimaciones por intervalo.



Media poblacional: σ conocida

Con objeto de obtener una estimación por intervalo para la media poblacional, se necesita la desviación estándar poblacional σ o la desviación estándar muestral s para calcular el margen de error. En la mayor parte de los casos, no se conoce σ y para calcular el margen de error se emplea s. Sin embargo, en algunas ocasiones, se cuenta con una gran cantidad de datos anteriores (históricos) que se pueden usar para calcular la desviación estándar poblacional antes de tomar la muestra. También, en aplicaciones al control de calidad, en las que se supone que el proceso se desarrolla correctamente "en control", se considera que se conoce la desviación estándar. A tales casos se les conoce como casos **de σ conocida**. En esta sección se presenta un caso en el que es razonable considerar que se conoce σ y se muestra cómo obtener una estimación por intervalo.

Cada semana, la empresa Lloyd's Department Store selecciona una muestra aleatoria simple de 100 clientes con objeto de conseguir información acerca de la cantidad que gastan en cada visita a la tienda. Si x representa la cantidad gastada en cada visita a la tienda, la media muestral $\bar{x}$ es una estimación puntual de μ, la cantidad media gastada en cada visita a la tienda por la población formada por los clientes de Lloyd's Department Store. Lloyd's ha estado realizando estos estudios semanales durante varios años. Con base en sus datos anteriores, Lloyd´s supone que el valor conocido de la desviación estándar poblacional es $\sigma = \$20$. Los datos anteriores (históricos) indican también que la población tiene una distribución normal.



Lloyd's

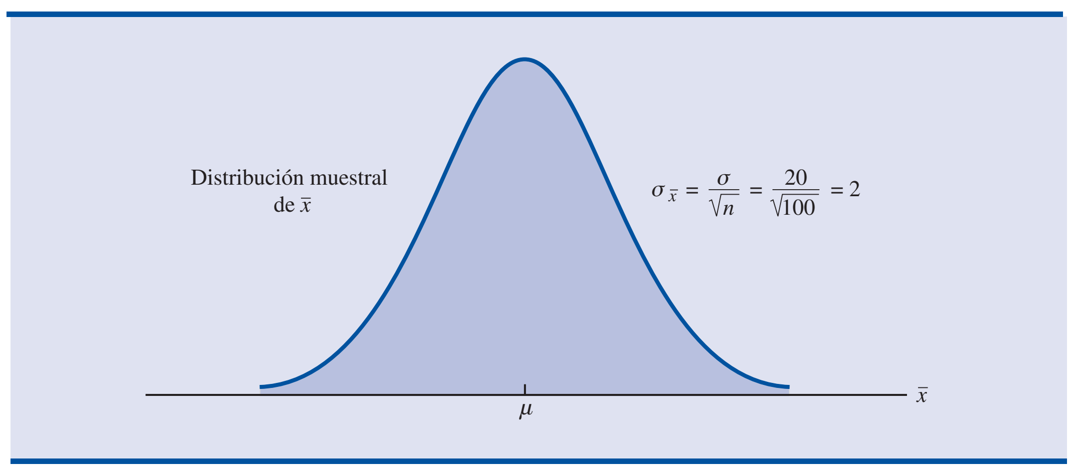
En la última semana, en su estudio de 100 clientes ($n = 100$), Lloyd's obtuvo como media muestral $\bar{x} = \$82$. La media muestral de la cantidad gastada permite una estimación puntual de la media poblacional de la cantidad gastada en cada visita, μ. A continuación se muestra cómo calcular un margen de error para esta estimación y cómo obtener una estimación por intervalo para la media poblacional.

Margen de error y estimación por intervalo

En el capítulo 7 se mostró que la distribución muestral de $\bar{x}$ sirve para calcular la probabilidad de que $\bar{x}$ esté dentro de una distancia dada de μ. En el ejemplo de Lloyd´s, los datos históricos muestran que la población constituida por las cantidades gastadas está distribuida normalmente y que su desviación estándar es $\sigma = 20$. De manera que, usando lo aprendido en el capítulo 7, se puede concluir que la distribución muestral de $\bar{x}$ es una distribución normal con error estándar de $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 20/\sqrt{100} = 2$. En la figura 8.1 se muestra esta distribución muestral.* Puesto que la

* Se aprovecha que las cantidades gastadas tienen una distribución normal para concluir que la distribución muestral de $\bar{x}$ tiene una distribución normal. Si la población no tuviera una distribución normal, de acuerdo con el teorema del límite central y dado que el tamaño de la muestra $n = 100$ se puede concluir que la distribución muestral de $\bar{x}$ es aproximadamente normal. De cualquier manera, la distribución muestral de $\bar{x}$ es como se muestra en la figura 8.1.

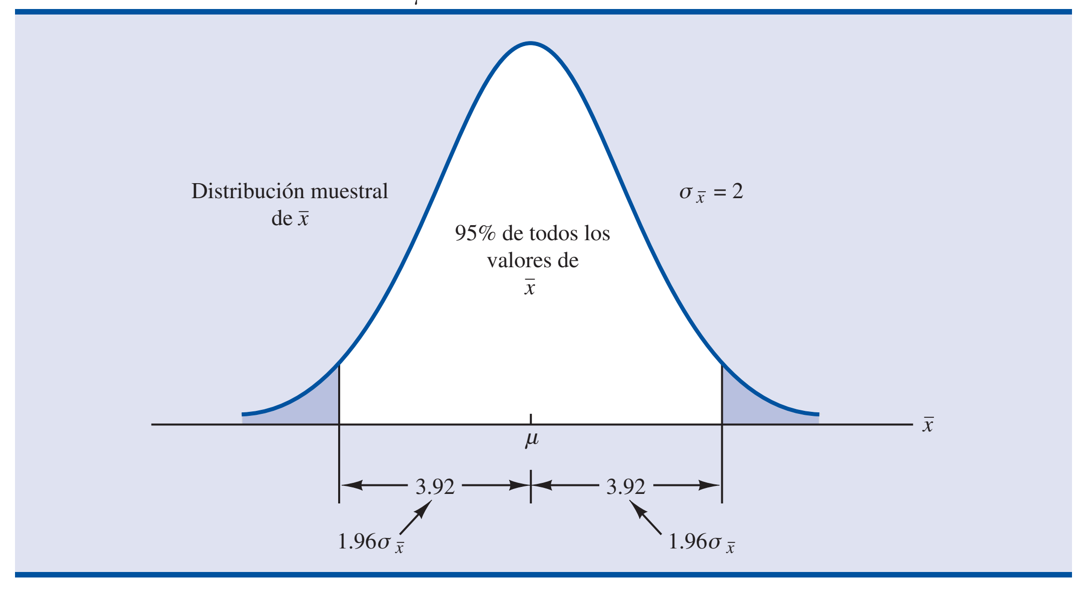
FIGURA 8.1 DISTRIBUCIÓN MUESTRAL DE LA MEDIA MUESTRAL DE LAS CANTIDADES GASTADAS EN MUESTRAS ALEATORIAS SIMPLES DE 100 CLIENTES



distribución muestral indica cómo están distribuidos los valores de $\bar{x}$ en torno a la media poblacional μ, la distribución muestral de $\bar{x}$ proporciona información acerca de la posible diferencia entre $\bar{x}$ y μ.

En la tabla de probabilidad normal estándar se encuentra que 95% de los valores de cualquier variable aleatoria distribuida normalmente aparecen dentro de $\pm$ 1.96 desviaciones estándar de la media. Por tanto, si la distribución muestral se encuentra distribuida normalmente, 95% de los valores de $\bar{x}$ deben estar dentro de $\pm 1.96\sigma_{\bar{x}}$ de la media μ. En el ejemplo de Lloyd´s, se sabe que la distribución muestral de $\bar{x}$ está distribuida normalmente y que el error estándar es $\sigma_{\bar{x}} = 2$. Como $\pm 1.96\sigma_{\bar{x}} = 1.96(2) = 3.92$, se puede concluir que 95% de los valores de $\bar{x}$ obtenidos usando muestras de $n = 100$ estarán dentro de ± 3.92 de la media poblacional. Véase figura 8.2.

FIGURA 8.2 DISTRIBUCIÓN MUESTRAL DE $\bar{x}$ EN LA QUE SE MUESTRA LA LOCALIZACIÓN DE LAS MEDIAS MUESTRALES QUE SE ENCUENTRAN DENTRO DE 3.92 DE μ.

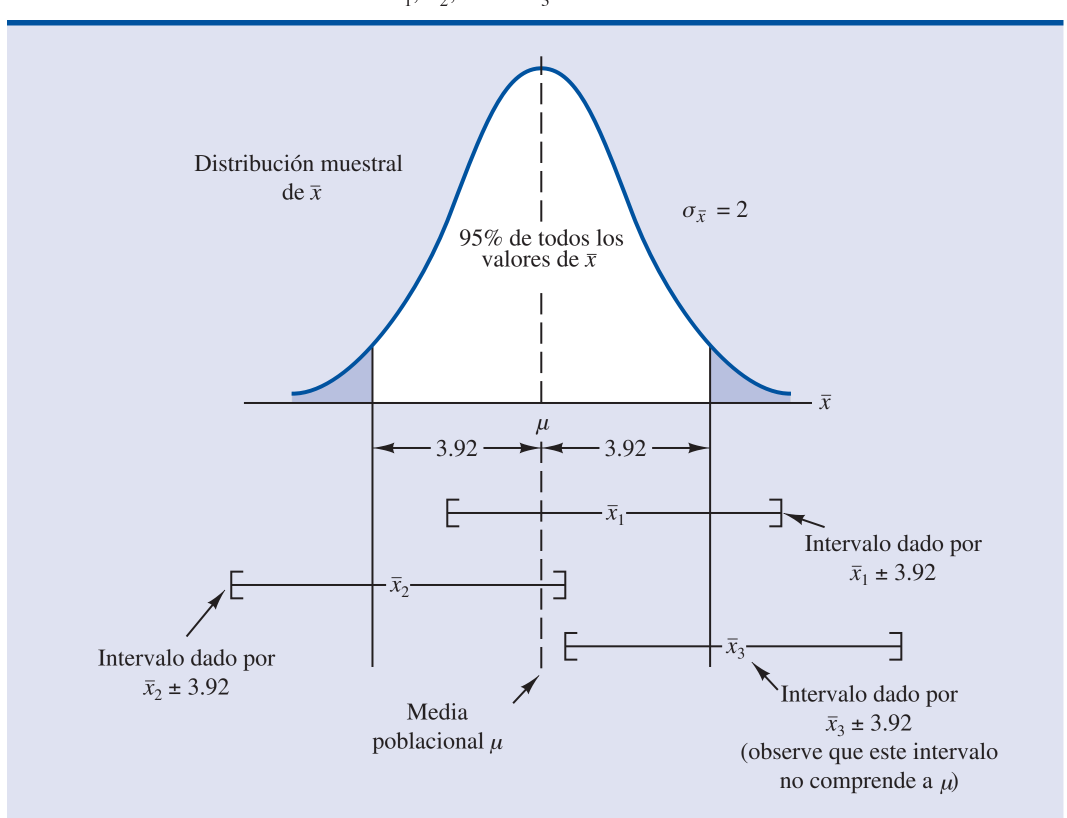


En la introducción a este capítulo se dijo que la fórmula general para estimar un intervalo de la media poblacional μ es $\bar{x} \pm$ margen de error. En el ejemplo de Lloyd's, suponga que se establece 3.92 como margen de error y se calcula una estimación por intervalo para μ usando $\bar{x} \pm$ 3.92. Para ver cómo se interpreta dicha estimación por intervalo, considere los valores de $\bar{x}$ que podrían obtenerse si se tomaran tres muestras aleatorias simples *diferentes,* cada una de 100 clientes de Lloyd´s. La primera media muestral puede que dé el valor $\bar{x}_1$ de la figura 8.3. En este caso, como se ve en la figura, el intervalo que se obtiene restando 3.92 de $\bar{x}_1$ y sumando 3.92 a $\bar{x}_1$ abarca a la media poblacional μ. Ahora razone lo que pasa si la segunda media muestral resulta tener el valor $\bar{x}_2$ que se muestra en la figura 8.3. Aunque esta media muestral difiere de la primera media muestral, el intervalo obtenido restando 3.92 de $\bar{x}_2$ y sumando 3.92 a $\bar{x}_2$ también abarca a la media poblacional μ. Pero, considere lo que pasa si la tercera media muestral resulta tener el valor $\bar{x}_3$ que se muestra en la figura 8.3. En este caso el intervalo obtenido restando 3.92 de $\bar{x}_3$ y sumando 3.92 a $\bar{x}_3$ no abarca a la media poblacional μ. Como $\bar{x}_3$ cae en la cola superior de la distribución muestral y dista más de 3.92 de μ, restando y sumando 3.92 de $\bar{x}_3$ se obtiene un intervalo que no abarca a μ.

Con cualquier media muestral $\bar{x}$ que se encuentre dentro de la región sombreada en la figura 8.3, se obtendrá un intervalo que contenga a la media poblacional μ. Como 95% de todas las posibles medias muestrales se encuentran en la región sombreada, 95% de todos los intervalos que se obtengan restando 3.92 de $\bar{x}$ y sumando 3.92 a $\bar{x}$ abarcarán a la media poblacional μ.

Recuerde que en la última semana el equipo encargado del aseguramiento de la calidad de Lloyd´s encuestó a 100 clientes y obtuvo una media muestral de $\bar{x}$ = 82. Usando $\bar{x} \pm$ 3.92 para

FIGURA 8.3 INTERVALOS OBTENIDOS A PARTIR DE ALGUNAS MEDIAS MUESTRALES LOCALIZADAS EN $\bar{x}_1$, $\bar{x}_2$, Y EN $\bar{x}_3$



Esto aclara por qué se le llama intervalo de confianza de 95%.

obtener la estimación por intervalo, se obtiene 82 ± 3.92. Por tanto, la estimación por intervalo que se basa en los datos de la última semana es el intervalo que va de $82 - 3.92 = 78.08$ a $82 + 3.92 = 85.92$. Como 95% de todos los intervalos construidos usando $\bar{x} \pm 3.92$ contendrán a la media poblacional se tiene 95% de confianza de que el intervalo 78.08 a 85.92 contenga a la media poblacional μ. Entonces dicho intervalo tiene un **nivel de confianza** de 95%. Al valor 0.95 se le conoce como **coeficiente de confianza** y al intervalo 78.08 a 85.92 como el **intervalo de confianza de 95%**.

Como el margen de error está dado por $z_{\alpha/2}(\sigma/\sqrt{n})$, la fórmula general de una estimación por intervalo de la media poblacional con σ conocida, es la siguiente.

ESTIMACIÓN POR INTERVALO DE UNA MEDIA POBLACIONAL: σ CONOCIDA

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \textbf{(8.1)}$$

donde $(1 - \alpha)$ es el coeficiente de confianza y $z_{\alpha/2}$ es el valor de z que proporciona un área $\alpha/2$ en la cola superior de la distribución de probabilidad normal estándar.

En el ejemplo de Lloyd´s, mediante la expresión (8.1), se construye un intervalo de confianza de 95% con un coeficiente de confianza $(1 - \alpha) = 0.95$ y, por tanto, a = 0.05. En la tabla de la distribución normal estándar aparece que un área de $\alpha/2 = 0.05/2 = 0.025$ en la cola superior corresponde a $z_{0.025} = 1.96$. Como en el ejemplo de Lloyd´s, la media muestral es $\bar{x} = 82$, $\sigma = 20$ y el tamaño de la muestra es $n = 100$, se obtiene

$$82 \pm 1.96 \frac{20}{\sqrt{100}}$$
$$82 \pm 3.92$$

Por tanto, empleando la expresión (8.1), el margen de error es 3.92 y el intervalo de confianza de 95% va de $82 - 3.92 = 78.08$ a $82 + 3.92 = 85.92$.

Aunque a menudo se usa un nivel de confianza de 95%, también suelen usarse otros niveles de confianza como 90% y 99%. En la tabla 8.1 se muestran los valores de $z_{\alpha/2}$ correspondientes a los niveles de confianza más utilizados. A partir de estos valores y de la expresión (8.1), el intervalo de confianza de 90% en el ejemplo de Lloyd´s es

$$82 \pm 1.645 \frac{20}{\sqrt{100}}$$
$$82 \pm 3.29$$

TABLA 8.1 VALORES DE $z_{\alpha/2}$ PARA LOS NIVELES DE CONFIANZA MÁS USADOS

Nivel de confianza	α	$\alpha/2$	$z_{\alpha/2}$
90%	.10	.05	1.645
95%	.05	.025	1.960
99%	.01	.005	2.576

Por tanto, para 90% de confianza, el margen de error es 3.29 y el intervalo de confianza es 82 − 3.29 = 78.71 a 82 + 3.29 = 85.29. De manera similar, el intervalo de confianza de 99% es

$$82 \pm 2.576 \frac{20}{\sqrt{100}}$$

$$82 \pm 5.15$$

Entonces, para 99% de confianza el margen de error es 5.15 y el intervalo de confianza es 82 − 5.15 = 76.85 a 82 + 5.15 = 87.15.

Al comparar los resultados para los niveles de 90%, 95% y 99% de confianza, es claro que para tener mayor grado de confianza, el margen de error y con esto la amplitud del intervalo de confianza debe ser mayor.

Recomendación práctica

Si la población tiene una distribución normal, el intervalo de confianza que se obtiene con la expresión (8.1) es exacto. En otras palabras, si la expresión (8.1) se usa repetidas veces para generar intervalos de 95% de confianza, exactamente 95% de los intervalos generados contendrán la media poblacional. Si la población no tiene una distribución normal, el intervalo de confianza obtenido con la expresión (8.1) será aproximado. En este caso la calidad de la aproximación depende tanto de la distribución de la población como del tamaño de la muestra.

En la mayoría de las aplicaciones, cuando se usa la expresión (8.1), un tamaño de muestra $n \geq 30$ es adecuado para obtener una estimación por intervalo de la media poblacional. Si la población no está distribuida normalmente, pero es más o menos simétrica, tamaños de muestra hasta de 15 puede esperarse que proporcionen una buena aproximación del intervalo de confianza. Con tamaños de muestra menores, la expresión (8.1) sólo se debe usar si el analista cree, o está dispuesto a suponer, que la distribución de la población es cuando menos aproximadamente normal.

NOTAS Y COMENTARIOS

1. El procedimiento de estimación por intervalo visto en esta sección se basa en la suposición de que la desviación estándar poblacional σ sea conocida. Decir que σ es conocida significa que se cuenta con datos históricos o con otra información que permita obtener una buena estimación de la desviación estándar antes de tomar la muestra que se usará para obtener la estimación de la media poblacional. De manera que, técnicamente esto no significa que σ se conozca con seguridad. Lo que significa es que sólo se obtuvo una buena estimación de la desviación estándar antes de tomar la muestra y que de esta manera no se usará la misma muestra para estimar tanto la media poblacional como la desviación estándar poblacional.

2. El tamaño de la muestra n aparece en el denominador de la expresión (8.1) para la estimación por intervalo. En consecuencia, si un determinado tamaño de muestra da un intervalo demasiado amplio, para que tenga utilidad práctica, se aumenta el tamaño de la muestra. Si n está en el denominador, con un tamaño de muestra mayor se obtendrá un margen de error menor, un intervalo más estrecho y mayor precisión. El procedimiento para determinar el tamaño de una muestra aleatoria simple que se necesita para obtener una determinada precisión se verá en la sección 8.3.

Ejercicios

Métodos

1. En una muestra aleatoria simple de 40 artículos la media muestral obtenida es 25. La desviación estándar poblacional es $\sigma = 5$.
   a. ¿Cuál es el error estándar de la media $\sigma_{\bar{x}}$?
   b. ¿Cuál es el margen de error para tener 95% de confianza?

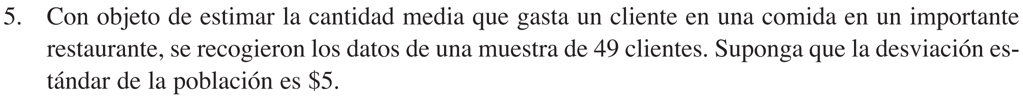


2. En una muestra aleatoria simple de 50 artículos de una población en la que $\sigma = 6$ la media muestral fue 32.
   a. Proporcione un intervalo de confianza de 90% para la media poblacional.
   b. Establezca un intervalo de confianza de 95% para la media poblacional.
   c. Proporcione un intervalo de confianza de 99% para la media poblacional.
3. En una muestra aleatoria simple de 60 artículos la media muestral fue 80. La desviación estándar poblacional es $\sigma = 15$.
   a. Calcule el intervalo de confianza de 95% para la media poblacional.
   b. Suponga que la misma media muestral se obtuvo de una muestra de 120 artículos. Dé el intervalo de confianza de 95% para la media poblacional.
4. Para la media poblacional se dio el siguiente intervalo de confianza de 95%, de 152 a 160. Si $\sigma = 15$, ¿cuál es el tamaño de la muestra que se usó en este estudio?



Aplicaciones





5. Con objeto de estimar la cantidad media que gasta un cliente en una comida en un importante restaurante, se recogieron los datos de una muestra de 49 clientes. Suponga que la desviación estándar de la población es \$5.
   a. ¿Cuál es el margen de error para 95% de confianza?
   b. Si la media poblacional es \$24.80, ¿cuál es el intervalo de confianza de 95% para la media poblacional?



6. Nielsen Media Research llevó a cabo un estudio para saber cuánto tiempo se veía televisión en los hogares, en el horario de 8:00 a 11:00 de la noche. Los datos que se encuentran en el archivo Nielsen del disco compacto son consistentes con los hallazgos reportados (*The World Almanac*, 2003). Con base en estudios anteriores, la desviación estándar poblacional se considera conocida y es $\sigma = 3.5$ horas. Dé una estimación mediante un intervalo de confianza de 95% para la media del tiempo que se ve televisión por semana en el horario de 8:00 a 11:00 de la noche.
7. En una investigación sobre los negocios pequeños que tienen un sitio en la Web se encontró que la cantidad promedio que se gasta en un sitio es \$11 500 por año. Dada una muestra de 60 negocios y una desviación estándar $\sigma = \$4000$, ¿cuál es el margen de error? Use 95% de confianza. ¿Qué recomendaría si el estudio requiere un margen de error de \$500?
8. The National Quality Research Center de la University of Michigan proporciona medidas trimestrales de las opiniones de los consumidores acerca de productos y servicios (*The Wall Street Journal*, 18 de febrero de 2003). En una encuesta sobre 10 restaurantes de comida rápida y pizza la media del índice de satisfacción de los consumidores fue 71. Datos anteriores indican que la desviación estándar ha sido relativamente estable y es $\sigma = 5$.
   a. ¿Qué debe estar dispuesto a suponer el investigador si desea un margen de error?
   b. Con 95% de confianza, ¿cuál es el margen de error?
   c. ¿Cuál es el margen de error si se desea 99% de confianza?
9. El puntuaje promedio en el examen de admisión de los estudiantes que ingresaron a la escuela de negocios fue 3.37 (*Best Graduate Schools*, *U.S. News and World Report*, 2001). Suponga que dicha estimación se basó en una muestra de 120 estudiantes. De acuerdo con datos anteriores se admite que se conoce la desviación estándar y que es $\sigma = 0.28$. ¿Cuál es la estimación mediante un intervalo de confianza de 95% para la media del puntuaje promedio de los alumnos que ingresaron a la escuela de negocios?
10. La revista *Playbill* reportó que el ingreso familiar anual medio de sus suscriptores es \$119 155 (*Playbill*, enero de 2006). Suponga que la estimación del ingreso familiar anual medio está basada en una muestra de 80 familias y que por datos de estudios anteriores la desviación estándar poblacional es conocida y es $\sigma = \$30\,000$.

a. Desarrolle un intervalo de estimación de 90% de confianza para la media poblacional.
b. Dé un intervalo de estimación de 95% de confianza para la media poblacional.
c. Dé un intervalo de estimación de 99% de confianza para la media poblacional.
d. ¿Qué le pasa a la amplitud del intervalo de confianza a medida que el nivel de confianza aumenta? ¿Parece esto razonable? Explique.



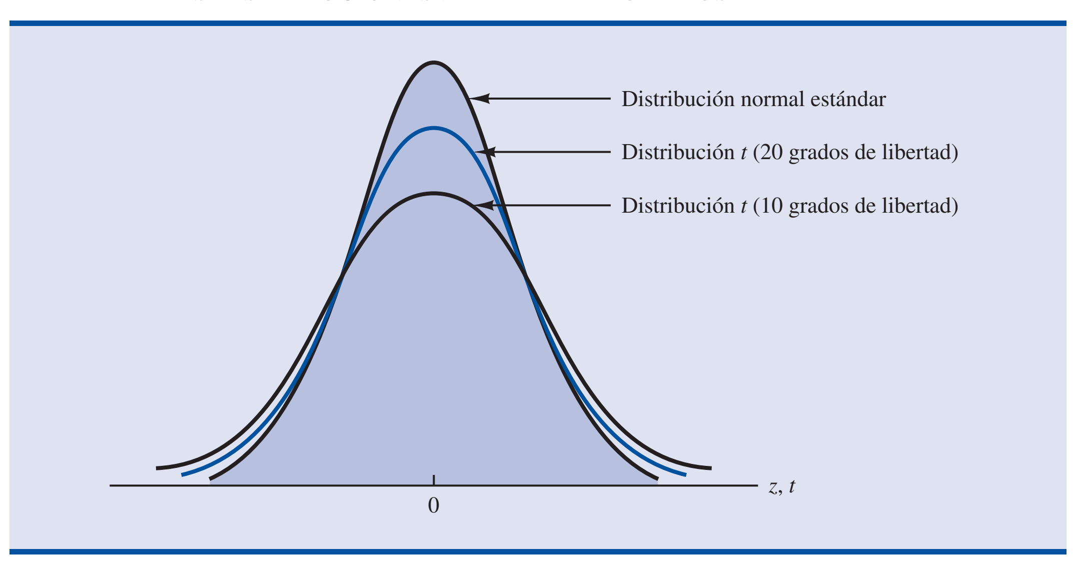
8.2 Media poblacional: σ desconocida

William Sealy Gosset, quien publicaba bajo el seudónimo "Student" estableció la distribución t. Gosset, que había estudiado matemáticas en Oxford, trabajaba para Guinness Brewery en Dublín, Irlanda. Elaboró la distribución t cuando trabajaba con materiales a pequeña escala y hacía experimentos de temperatura.

Cuando se calcula un intervalo de confianza para la media poblacional, suele no contarse con una buena estimación de la desviación estándar poblacional. En tales casos se usa la misma muestra para estimar μ y σ. Esta situación es el caso que se conoce como **σ desconocida**. Cuando se usa s para estimar σ, el margen de error y la estimación por intervalo de la media poblacional se basan en una distribución de probabilidad conocida como **distribución t**. Aunque la elaboración matemática de la distribución t parte de la suposición de que la población de la que se muestrea tiene una distribución normal, las investigaciones han demostrado que la distribución t se aplica en muchas situaciones en que la población se desvía significantemente de una población normal. Más adelante, en esta misma sección, se proporcionan lineamientos para usar la distribución t cuando la población no está distribuida normalmente.

La distribución t es una familia de distribuciones de probabilidad similares; cada distribución t depende de un parámetro conocido como **grados de libertad**. La distribución t para un grado de libertad es única, como lo es la distribución t para dos grados de libertad, para tres grados de libertad, etc. A medida que el número de grados de libertad aumenta, la diferencia entre la distribución t y la distribución normal estándar se va reduciendo. En la figura 8.4 se muestran las distribuciones t para 10 y 20 grados de libertad y su relación con la distribución de probabilidad normal estándar. Observe que una distribución t para más grados de libertad exhibe menos va-

FIGURA 8.4 COMPARACIÓN DE LA DISTRIBUCIÓN NORMAL ESTÁNDAR CON LAS DISTRIBUCIONES t PARA 10 Y 20 GRADOS DE LIBERTAD



riabilidad y un mayor parecido con la distribución normal estándar, también que la media de toda distribución t es cero.

Para indicar el área en la cola superior de la distribución t, a la t se le pone un subíndice. Por ejemplo, así como se usó $z_{0.025}$ para indicar el valor de z que deja en la cola superior de la distribución normal estándar un área de 0.025, se usará $t_{0.025}$ para indicar el valor de t que deja en la cola superior de la distribución t un área de 0.025. En general se usará la notación $t_{\alpha/2}$ para representar el valor de t que deja un área de $\alpha/2$ en la cola superior de la distribución t. Véase figura 8.5.

La tabla 2 del apéndice B contiene una tabla con la distribución t. En la tabla 8.2 se muestra parte de esa tabla. Cada renglón de la tabla corresponde a una distribución t distinta con los grados de libertad que se indican. Por ejemplo, en la distribución t con 9 grados de libertad, $t_{0.025} = 2.262$. De manera similar en la distribución t con 60 grados de libertad, $t_{0.025} = 2.000$. A medida que los grados de libertad aumentan, $t_{0.025}$ se aproxima a $z_{0.025} = 1.96$. En efecto, el valor z de la distribución normal estándar se encuentra en el renglón correspondiente a infinitos grados de libertad (cuyo encabezado es ∞) de la tabla de las distribuciones t. Si los grados de libertad son más de 100, se puede usar el renglón correspondiente a infinitos grados de libertad para aproximar el verdadero valor de t; en otras palabras, para más de 100 grados de libertad, el valor z normal estándar proporciona una buena aproximación al valor t.

A medida que los grados de libertad aumentan, la distribución t se aproxima más a la distribución normal estándar.

Margen de error en estimación por intervalo

En la sección 8.1 se mostró que la estimación por intervalo de la media poblacional cuando σ es conocida es

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Para calcular una estimación por intervalo para μ en el caso en que no se conoce σ se usa la desviación estándar muestral s para estimar σ, y $z_{\alpha/2}$ se sustituye por el valor $t_{a/2}$ de la distribución t.

FIGURA 8.5 ÁREA DE DISTRIBUCIÓN t CON UN ÁREA O PROBABILIDAD $\alpha/2$ EN LA COLA SUPERIOR

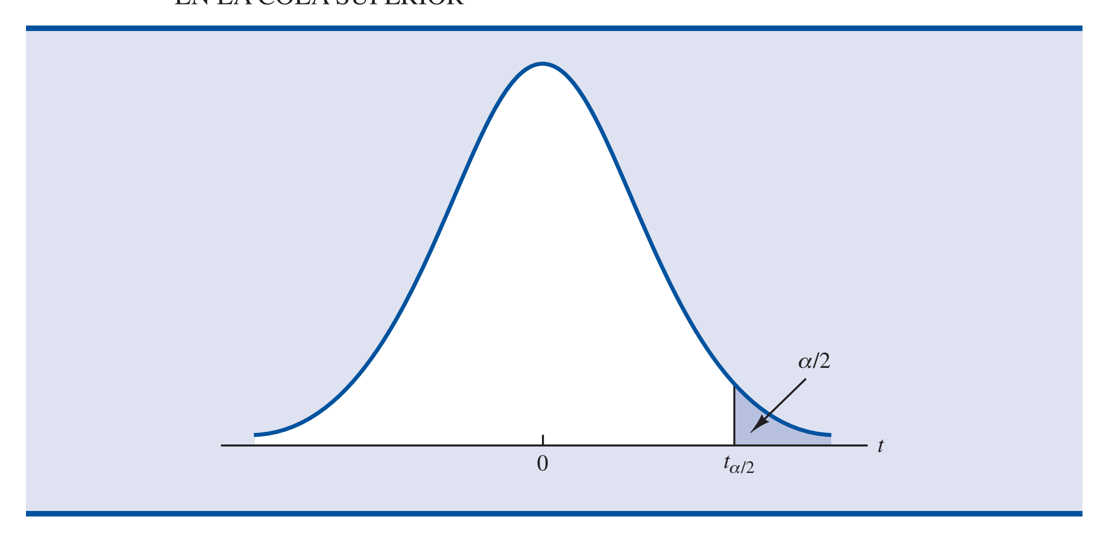
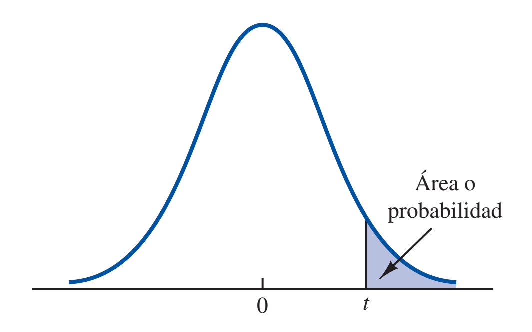


TABLA 8.2 ALGUNOS VALORES DE LA TABLA DE LA DISTRIBUCIÓN t*



Grados de libertad	Área en la cola superior					
	0.20	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	0.978	1.638	2.353	3.182	4.541	5.841
4	0.941	1.533	2.132	2.776	3.747	4.604
5	0.920	1.476	2.015	2.571	3.365	4.032
6	0.906	1.440	1.943	2.447	3.143	3.707
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1.397	1.860	2.306	2.896	3.355
9	0.883	1.383	1.833	2.262	2.821	3.250
60	0.848	1.296	1.671	2.000	2.390	2.660
61	0.848	1.296	1.670	2.000	2.389	2.659
62	0.847	1.295	1.670	1.999	2.388	2.657
63	0.847	1.295	1.669	1.998	2.387	2.656
64	0.847	1.295	1.669	1.998	2.386	2.655
65	0.847	1.295	1.669	1.997	2.385	2.654
66	0.847	1.295	1.668	1.997	2.384	2.652
67	0.847	1.294	1.668	1.996	2.383	2.651
68	0.847	1.294	1.668	1.995	2.382	2.650
69	0.847	1.294	1.667	1.995	2.382	2.649
90	0.846	1.291	1.662	1.987	2.368	2.632
91	0.846	1.291	1.662	1.986	2.368	2.631
92	0.846	1.291	1.662	1.986	2.368	2.630
93	0.846	1.291	1.661	1.986	2.367	2.630
94	0.845	1.291	1.661	1.986	2.367	2.629
95	0.845	1.291	1.661	1.985	2.366	2.629
96	0.845	1.290	1.661	1.985	2.366	2.628
97	0.845	1.290	1.661	1.985	2.365	2.627
98	0.845	1.290	1.661	1.984	2.365	2.627
99	0.845	1.290	1.660	1.984	2.364	2.626
100	0.845	1.290	1.660	1.984	2.364	2.626
∞	0.842	1.282	1.645	1.960	2.326	2.576

**Nota:* Una tabla más extensa se encuentra en la tabla 2 del apéndice B.

El margen de error está dado, entonces, por $t_{\alpha/2} s/\sqrt{n}$. Con este margen de error, la expresión general para una estimación por intervalo de la media poblacional cuando σ no se conoce es la siguiente.

ESTIMACIÓN POR INTERVALO DE LA MEDIA POBLACIONAL: σ DESCONOCIDA

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \tag{8.2}$$

donde s es la desviación estándar muestral, $(1 - \alpha)$ es el coeficiente de confianza y $t_{\alpha/2}$ es el valor de t que proporciona un área de $\alpha/2$ en la cola superior de la distribución t para $n - 1$ grados de libertad.

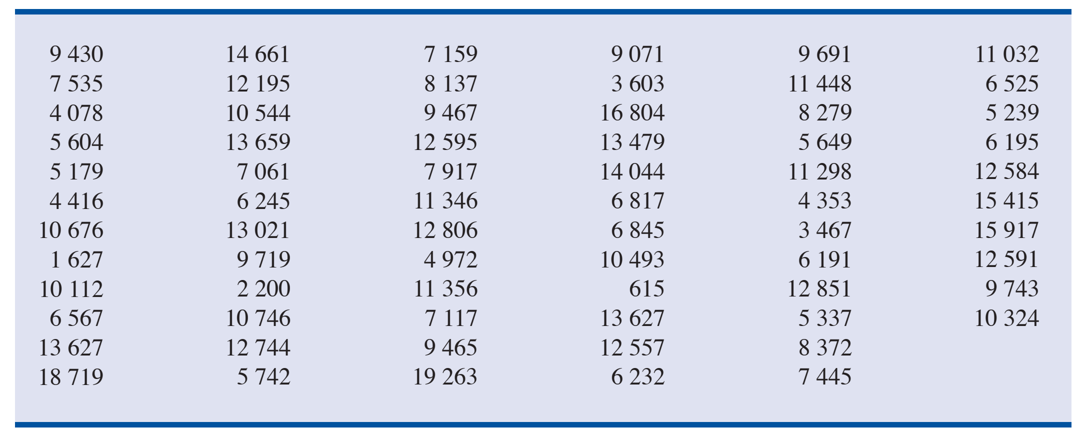
La razón de que el número de grados de libertad para el valor de t en la expresión (8.2) sea $n - 1$ se debe al uso de s como estimación de la desviación estándar poblacional σ. La expresión para calcular la desviación estándar muestral es

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}}$$

Los grados de libertad se refieren al número de valores independientes en el cálculo de $\Sigma(x_i - \bar{x})^2$. Los n valores en el cálculo de $\Sigma(x_i - \bar{x})^2$ son los siguientes: $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$. En la sección 3.2 se indicó que en cualquier conjunto de datos $\Sigma(x_i - \bar{x}) = 0$. Por tanto, únicamente $n - 1$ de los valores $x_i - \bar{x}$ son independientes; es decir, si se conocen $n - 1$ de estos valores, el valor restante puede determinarse exactamente usando el hecho de que los valores $x_i - \bar{x}$ deben sumar 0. Entonces, $n - 1$ es el número de grados de libertad en la suma $\Sigma(x_i - \bar{x})^2$ y de ahí el número de grados de libertad para la distribución t en la expresión (8.2).

Para ilustrar la estimación por intervalo en el caso σ desconocida, se verá un estudio realizado para estimar la media del adeudo en las tarjetas de crédito en la población de familias de Estados Unidos. En la tabla 8.3 se presentan los saldos en las tarjetas de crédito de una muestra de $n = 70$ familias. En esta ocasión no se cuenta con una estimación previa de la desviación estándar poblacional σ. De manera que los datos muestrales deberán usarse para estimar tanto la media poblacional como la desviación estándar poblacional. Con los datos de la tabla 8.3 se calcula

TABLA 8.3 SALDOS EN LAS TARJETAS DE CRÉDITO DE UNA MUESTRA DE 70 FAMILIAS



9 430	14 661	7 159	9 071	9 691	11 032
7 535	12 195	8 137	3 603	11 448	6 525
4 078	10 544	9 467	16 804	8 279	5 239
5 604	13 659	12 595	13 479	5 649	6 195
5 179	7 061	7 917	14 044	11 298	12 584
4 416	6 245	11 346	6 817	4 353	15 415
10 676	13 021	12 806	6 845	3 467	15 917
1 627	9 719	4 972	10 493	6 191	12 591
10 112	2 200	11 356	615	12 851	9 743
6 567	10 746	7 117	13 627	5 337	10 324
13 627	12 744	9 465	12 557	8 372	
18 719	5 742	19 263	6 232	7 445	



la media muestral $\bar{x} = \$9312$ y la desviación estándar muestral $s = \$4007$. Ahora se usa la tabla 8.2 para obtener el valor de $t_{0.025}$ correspondiente a 95% de confianza y $n - 1 = 69$ grados de libertad. El valor de t que se necesita está en el renglón correspondiente a 69 grados de libertad y en la columna correspondiente a 0.025 en la cola superior. El valor que se encuentra es $t_{.025} = 1.995$.

Con la expresión (8.2), para calcular la estimación por intervalo de la media poblacional de los saldos en las tarjetas de crédito, se tiene:

$$9312 \pm 1.995 \frac{4007}{\sqrt{70}}$$

$$9312 \pm 955$$

La estimación puntual de la media poblacional es \$9312, el margen de error es \$955 y el intervalo de confianza de 95% va de $9312 - 955 = \$8357$ a $9312 + 955 = \$10\,267$. En consecuencia, 95% de confianza de la media de los saldos en las tarjetas de crédito de la población de todas las familias está ente \$8357 y \$10 267.

En los apéndices 8.1 y 8.2 se describen los procedimientos para obtener un intervalo de confianza para la media poblacional usando Minitab y Excel. En la figura 8.6 se muestran los resultados para el estudio de los saldos en las tarjetas de crédito que da el procedimiento de Minitab para la estimación por intervalo. Con la muestra de 70 familias se obtiene una media muestral de \$9312 para los saldos en las tarjetas de crédito, una desviación estándar muestral de \$4007, una estimación del error estándar de la media de \$479 (valor redondeado) y un intervalo de confianza de 95% que va de \$8357 a \$10 267.

Recomendación práctica

Si la población tiene una distribución normal, el intervalo de confianza suministrado en la expresión (8.2) es exacto y se puede usar con cualquier tamaño de muestra. Si la población no sigue una distribución normal, el intervalo de confianza en la expresión (8.2) será aproximado. En este caso la calidad de la aproximación depende tanto de la distribución de la población como del tamaño de la muestra.

Cuando la distribución de la población es muy sesgada o cuando hay observaciones atípicas se necesitan muestras grandes.

En la mayoría de las aplicaciones un tamaño de muestra $n \geq 30$ es suficiente al usar la expresión (8.2), para obtener una estimación por intervalo de la media poblacional. Sin embargo, si la distribución de la población es muy sesgada o si hay observaciones atípicas, la mayoría de los especialistas en estadística recomienda un tamaño de muestra de 50 o más. Si la población no tiene una distribución normal pero es más o menos simétrica, ya con un tamaño de muestra de 15 puede esperarse una buena aproximación al intervalo de confianza. Con muestras más pequeñas la expresión (8.2) sólo debe usarse si el analista cree, o está dispuesto a suponer, que la distribución de la población es por lo menos aproximadamente normal.

Uso de una muestra pequeña

En el ejemplo siguiente se obtiene una estimación por intervalo para una media poblacional teniendo una muestra pequeña. Como ya se indicó, conocer la distribución de la población es importante para decidir si mediante una estimación por intervalo se obtendrán resultados aceptables.

Scheer Industries está considerando un nuevo programa asistido por computadora con el fin de capacitar a los empleados de mantenimiento para realizar la reparación de las máquinas. Con

FIGURA 8.6 INTERVALO DE CONFIANZA DE MINITAB PARA EL ESTUDIO DE LOS SALDOS EN LAS TARJETAS DE CRÉDITO

Variable	N	Mean	StDev	SE Mean	95% CI
NewBalance	70	9312.00	4007.00	478.93	(8356.56, 10 267.44)

TABLA 8.4 DURACIÓN DE LA CAPACITACIÓN, EN DÍAS, EN LA MUESTRA DE 20 EMPLEADOS DE SCHEER INDUSTRIES



Scheer

52	59	54	42
44	50	42	48
55	54	60	55
44	62	62	57
45	46	43	56

objeto de evaluar este programa, el director de manufactura solicita una estimación de la media poblacional del tiempo requerido para que los empleados de mantenimiento completen la capacitación asistida por computadora.

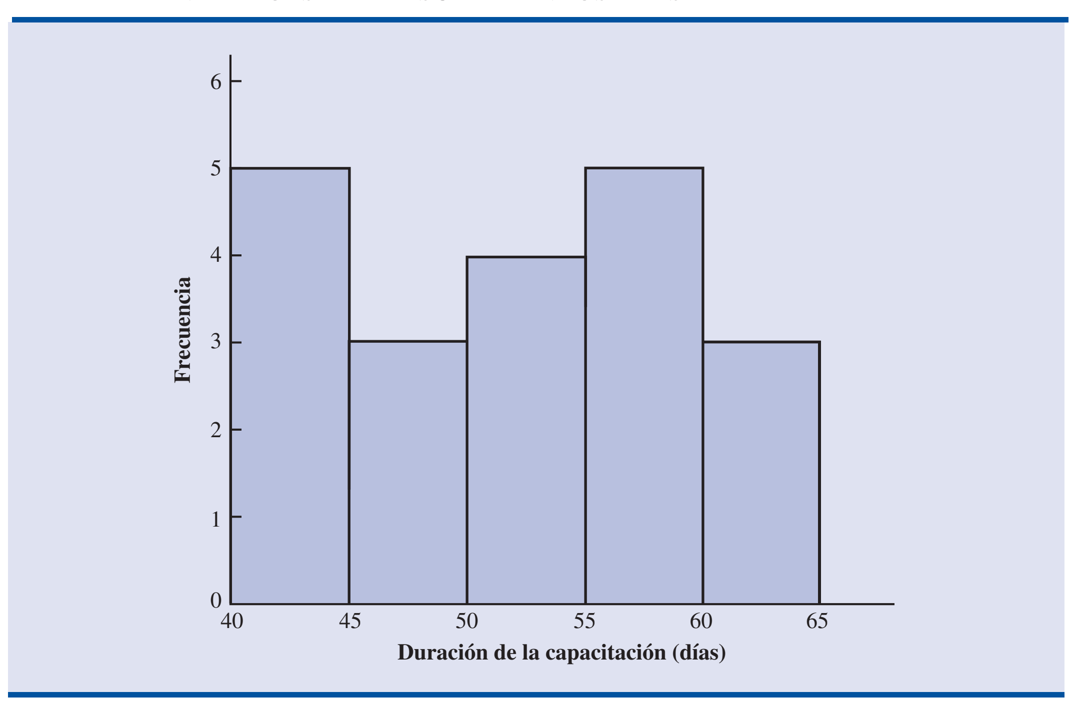
Considere una muestra de 20 empleados que siguen el programa de capacitación. En la tabla 8.4 se muestran los datos del tiempo, en días, que necesitó cada uno de los empleados para el programa de capacitación. En la figura 8.7 aparece un histograma de los datos. De acuerdo al histograma, ¿qué se puede decir de la distribución de estos datos? Primero, de acuerdo con los datos muestrales, no es posible concluir que la población sea normal, si bien no se observan evidencias de sesgo o de observaciones atípicas. Por tanto, mediante los lineamientos de la subsección anterior, se concluye que una estimación por intervalo basada en la distribución t parece ser aceptable para esta muestra de 20 empleados.

A continuación se calcula la media muestral y la desviación estándar muestral.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1030}{20} = 51.5 \text{ días}$$

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{889}{20 - 1}} = 6.84 \text{ días}$$

FIGURA 8.7 HISTOGRAMA SOBRE LA DURACIÓN DE LA CAPACITACIÓN EN LA MUESTRA DE SCHEER INDUSTRIES



Para dar un intervalo de confianza de 95%, se usa la tabla 2 del apéndice B y $n - 1 = 19$ grados de libertad y se obtiene $t_{0.025} = 2.093$. La expresión (8.2) suministra la estimación por intervalo de la media poblacional.

$$51.5 \pm 2.093\left(\frac{6.84}{\sqrt{20}}\right)$$

$$51.5 \pm 3.2$$

La estimación puntual de la media poblacional es 51.5 días. El margen de error es 3.2 días y el intervalo de confianza de 95% va de $51.5 - 3.2 = 48.3$ días a $51.5 + 3.2 = 54.7$ días.

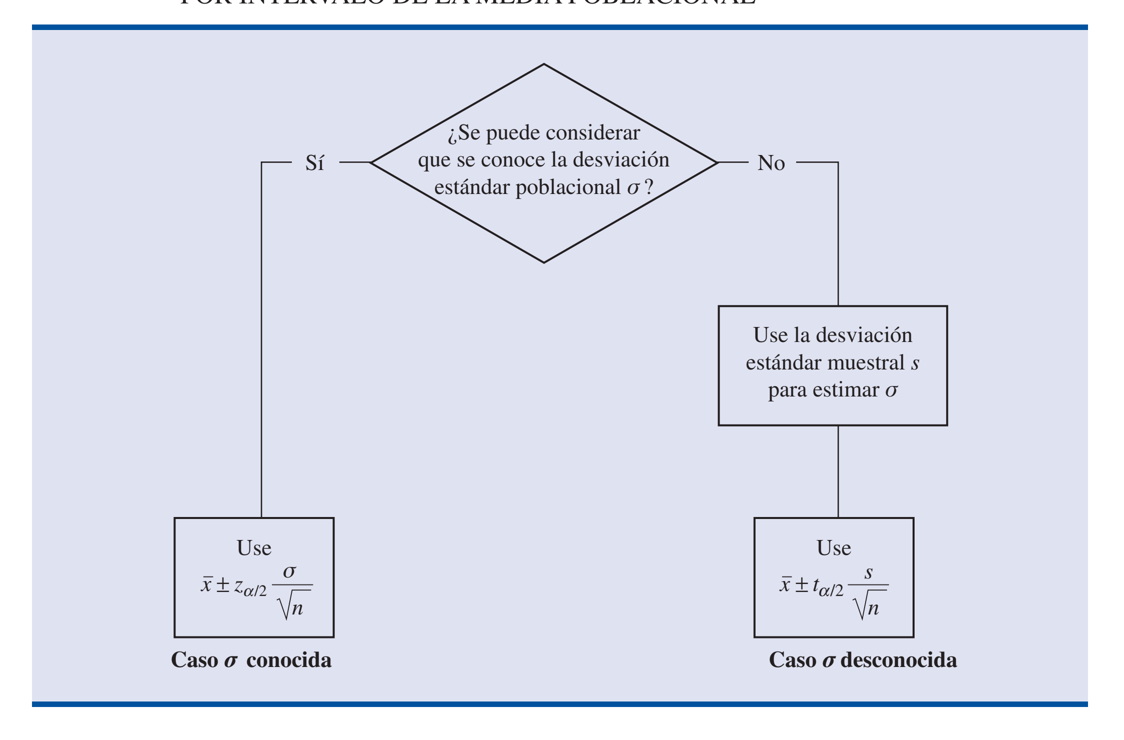
Usar un histograma de los datos muestrales para tener información acerca de la distribución de la población no es siempre concluyente, pero en muchos casos es la única información disponible. El histograma, junto con la opinión del analista, suele usarse para decidir si es adecuado usar la expresión (8.2) para obtener una estimación por intervalo.

Resumen de los procedimientos de estimación por intervalo

Se presentaron dos métodos para calcular una estimación por intervalo para la media poblacional. En el caso en que σ es conocida, en la expresión (8.1) se usan σ y la distribución normal estándar para calcular el margen de error y dar la estimación por intervalo. En el caso en que σ no es conocida, en la expresión (8.2) se usan la desviación estándar muestral s y la distribución t para calcular el margen de error y dar una estimación por intervalo.

En la figura 8.8 se presenta un resumen de los procedimientos para la estimación por intervalo de los dos casos. En la mayoría de las aplicaciones un tamaño de muestra $n \geq 30$ es adecuado. Sin embargo, si la población tiene distribución normal o aproximadamente normal, se pueden usar tamaños de muestra menores. En caso de que no se conozca σ y si la distribución de la población es muy sesgada o existen observaciones atípicas, se recomienda que el tamaño de la muestra sea $n \geq 50$.

FIGURA 8.8 RESUMEN DE LOS PROCEDIMIENTOS PARA LA ESTIMACIÓN POR INTERVALO DE LA MEDIA POBLACIONAL



NOTAS Y COMENTARIOS

1. Cuando se conoce σ, el margen de error, $z_{\alpha/2}(\sigma/\sqrt{n})$, es fijo y es el mismo para todas las muestras de tamaño n. Cuando σ no se conoce, el margen de error, $t_{\alpha/2}(s/\sqrt{n})$, varía de una muestra a otra. Esta variación se debe a que la desviación estándar muestral s varía con la muestra que se seleccione. Si s es grande, se obtiene un margen de error grande, mientras que si s es pequeña, se obtiene un margen de error pequeño.

2. ¿Qué pasa con las estimaciones por intervalo cuando la población es sesgada? Considere una población sesgada a la derecha, en la cual los datos con valores grandes jalan la distribución hacia la derecha. Cuando existe un sesgo así, hay una correlación positiva entre la media muestral $\bar{x}$ y la desviación estándar muestral s. Valores mayores de s tienden a corresponder a valores mayores de $\bar{x}$. De esta manera, cuando $\bar{x}$ es mayor que la media poblacional, s tiende a ser mayor que σ. Este sesgo hace que el margen de error, $t_{\alpha/2}(s/\sqrt{n})$, sea mayor de lo que sería si se conociera σ. Un intervalo de confianza con un margen de error mayor tenderá a incluir con más frecuencia a la media poblacional μ que si se usara el verdadero valor σ. Pero cuando $\bar{x}$ es menor que la media poblacional, la correlación entre $\bar{x}$ y s hace que el margen de error sea menor. En este caso, dichos intervalos de confianza con menor margen de error incluirán a la media poblacional menos veces que si se conociera y se usara σ. Por esta razón se recomienda usar tamaños de muestra más grandes cuando la distribución de la población es muy sesgada.

Ejercicios

Métodos

11. En la distribución t con 16 grados de libertad, encuentre el área, o la probabilidad, de cada una de las regiones siguientes:
    a. A la derecha de 2.120
    b. A la izquierda de 1.337
    c. A la izquierda de −1.746
    d. A la derecha de 2.583
    e. Entre −2.120 y 2.120
    f. Entre −1.746 y 1.746

12. Encuentre los valores de t para las situaciones siguientes.
    a. Un área de 0.025 en la cola superior, con 12 grados de libertad
    b. Un área de 0.05 en la cola inferior, con 50 grados de libertad
    c. Un área de 0.01 en la cola superior, con 30 grados de libertad
    d. Entre los que queda 90% del área, con 25 grados de libertad
    e. Entre los que queda 95% del área, con 45 grados de libertad



13. Los datos muestrales siguientes provienen de una población normal: 10, 8, 12, 15, 13, 11, 6, 5.
    a. ¿Cuál es la estimación puntual de la media poblacional?
    b. ¿Cuál es la estimación puntual de la desviación estándar poblacional?
    c. Con 95% de confianza, ¿cuál es el margen de error para la estimación de la media poblacional?
    d. ¿Cuál es el intervalo de confianza de 95% para la media poblacional?

14. En una muestra aleatoria simple con $n = 54$ la media muestral fue 22.5 y la desviación estándar muestral 4.4.
    a. Encuentre un intervalo de confianza de 90% para la media poblacional.
    b. Dé un intervalo de confianza de 95% para la media poblacional.
    c. Dé un intervalo de confianza de 99% para la media poblacional.
    d. ¿Qué pasa con el margen de error y con el intervalo de confianza a medida que aumenta el nivel de confianza?

Aplicaciones



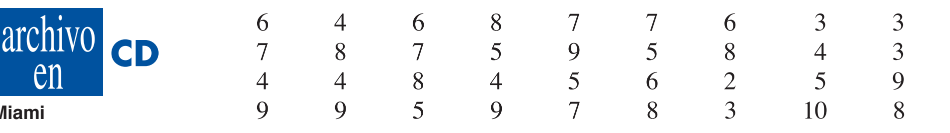
15. Los agentes de ventas de una empresa presentan un informe semanal que enumera los clientes contactados durante la semana. En una muestra de 65 informes semanales la media muestral es 19.5 clientes por semana. La desviación estándar es 5.2. Dé intervalos de confianza de 90% y 95% para la media poblacional del número de clientes contactados semanalmente por el personal de ventas.

16. El número medio de horas de vuelo de los pilotos de Continental Airlines es 49 horas por mes (*The Wall Street Journal*, 25 de febrero de 2003). Suponga que esta media se basó en las horas de vuelo de una muestra de 100 pilotos de esa empresa y que la desviación estándar muestral haya sido 8.5 horas.
    a. A 95% de confianza, ¿cuál es el margen de error?
    b. Dé el intervalo de estimación de 95% para la media poblacional de las horas de vuelo de los pilotos.
    c. La media en las horas de vuelo de los pilotos de United Airlines es 36 horas por mes. Use los resultados del inciso b para analizar la diferencia entre la cantidad de horas de vuelo de los pilotos en las dos líneas aéreas. *The Wall Street Journal* informa que United Airlines tiene el costo laboral más elevado de todas las aerolíneas. La información dada en estos ejercicios, ¿sirve para entender por qué se puede esperar que United Airlines tenga los costos más elevados?

17. La International Air Transport Association realiza encuestas entre los viajeros de negocios en las que se califica la calidad de los aeropuertos de salida internacional. La calificación máxima es 10. Se seleccionó una muestra aleatoria simple de 50 viajeros de negocios y a cada uno se le pidió su calificación para el aeropuerto internacional de Miami. Las calificaciones que dieron estos 50 viajeros fueron las que se muestran a continuación.



Miami



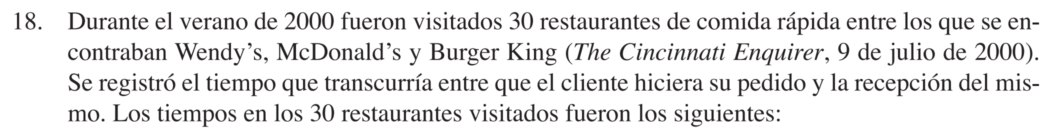
6	4	6	8	7	7	6	3	3	8	10	4	8
7	8	7	5	9	5	8	4	3	8	5	5	4
4	4	8	4	5	6	2	5	9	9	8	4	8
9	9	5	9	7	8	3	10	8	9	6		

Calcule el intervalo de confianza de 95% para la media poblacional de las calificaciones al aeropuerto de Miami.

18. Durante el verano de 2000 fueron visitados 30 restaurantes de comida rápida entre los que se encontraban Wendy's, McDonald's y Burger King (*The Cincinnati Enquirer*, 9 de julio de 2000). Se registró el tiempo que transcurría entre que el cliente hiciera su pedido y la recepción del mismo. Los tiempos en los 30 restaurantes visitados fueron los siguientes:



FastFood



0.9	1.0	1.2	2.2	1.9	3.6	2.8	5.2	1.8	2.1
6.8	1.3	3.0	4.5	2.8	2.3	2.7	5.7	4.8	3.5
2.6	3.3	5.0	4.0	7.2	9.1	2.8	3.6	7.3	9.0

    a. Dé una estimación puntual de la media poblacional.
    b. ¿Cuál es el margen de error con 95% de confianza?
    c. ¿Cuál es la estimación por intervalo de confianza de 95% para la media poblacional?
    d. Analice el sesgo que puede encontrarse en esta población. ¿Qué sugeriría para la repetición de este estudio?

19. En un estudio de National Retail Foundation se encontró que las familias estaban dispuestas a gastar en promedio $649 durante las vacaciones decembrinas (*The Wall Street Journal*, 2 de diciembre de 2002). Suponga que en el estudio participaron 600 familias y que la desviación estándar muestral fue $175.
    a. ¿Con 95% de confianza cuál es el margen de error?
    b. ¿Cuál es el intervalo de confianza de 95% para estimar la media poblacional?
    c. El año anterior, la media poblacional de gastos por familia fue $632. Analice la variación en el gasto en las vacaciones decembrinas en este periodo de un año.



20. ¿Los comerciales interrumpen constantemente su programa de televisión favorito? CNBC presentó datos estadísticos sobre la cantidad promedio de minutos de programa en media hora de transmisión (CNBC, 23 de febrero de 2006). Los datos siguientes (en minutos) son representativos de sus hallazgos.

21.06	22.24	20.62
21.66	21.23	23.86
23.82	20.30	21.52
21.52	21.91	23.14
20.02	22.20	21.20
22.37	22.19	22.34
23.36	23.44	

Suponga que la población es aproximadamente normal. Dé una estimación puntual y un intervalo de confianza de 95% para la cantidad media de minutos de programa en media hora de transmisión.



21. El consumo de las mujeres en edad de tomar bebidas alcohólicas ha aumentado en el Reino Unido, Estados Unidos y Europa (*The Wall Street Journal*, 15 de febrero de 2006). Datos (de consumo anual en litros) reportados por *The Wall Street Journal* hallados en una muestra de 20 mujeres jóvenes europeas son:

266	82	199	174	97
170	222	115	130	169
164	102	113	171	0
93	0	93	110	130

Si la población es más o menos simétrica, dé un intervalo de confianza de 95% para el consumo medio anual de bebidas alcohólicas entre las mujeres europeas jóvenes.

22. Las primeras semanas del 2004 fueron buenas para el mercado de acciones. En una muestra de 25 fondos abiertos se encontraron las siguientes ganancias obtenidas desde principio del año al 24 de enero del 2004 (*Barron's*, 19 de enero de 2004).



7.0	3.2	1.4	5.4	8.5
2.5	2.5	1.9	5.4	1.6
1.0	2.1	8.5	4.3	6.2
1.5	1.2	2.7	3.8	2.0
1.2	2.6	4.0	2.6	0.6

a. ¿Cuál es la estimación puntual de la media poblacional de las ganancias en fondos abiertos desde principio del año hasta esa fecha?
b. Puesto que la población tiene una distribución normal, calcule un intervalo de confianza de 95% para la media poblacional de las ganancias en fondos abiertos desde principio del año hasta esa fecha.



Determinación del tamaño de la muestra

El procedimiento que se presenta en esta sección se emplea para determinar el tamaño de muestra que se necesita para tener un determinado margen de error que se ha establecido antes de tomar la muestra.

En las recomendaciones prácticas de las dos secciones anteriores, se habló del papel del tamaño de la muestra para obtener una buena aproximación a los intervalos de confianza en los casos en que la población no tiene una distribución normal, ahora se enfoca la atención a otro aspecto relacionado con el tamaño de la muestra. Se describe cómo elegir un tamaño de muestra suficientemente grande para obtener un margen de error deseado. Para explicar esto, se vuelve al caso de la sección 8.1 en el que se tenía una σ conocida. Con la expresión (8.1), el intervalo de estimación está dado por

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

La cantidad $z_{\alpha/2}(\sigma/\sqrt{n})$ es el margen de error. De manera que, como se ve, $z_{\alpha/2}$, la desviación estándar poblacional σ, y el tamaño de la muestra n se combinan para determinar el margen de

error. Una vez que se selecciona el coeficiente de confianza $1 - \alpha$, se determina $z_{\alpha/2}$. Por tanto, si se tiene el valor de σ, es posible encontrar el tamaño de muestra n necesario para proporcionar cualquier margen de error deseado. A continuación se presenta la deducción de la fórmula que se usa para calcular el tamaño n de muestra deseado.

Sea $E =$ el margen de error deseado

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Despejando $\sqrt{n}$, se tiene

$$\sqrt{n} = \frac{z_{\alpha/2}\sigma}{E}$$

Al elevar al cuadrado ambos lados de esta ecuación, se obtiene la expresión siguiente para el tamaño de la muestra.

La ecuación (8.3) se usa para determinar el tamaño de muestra adecuado. Sin embargo, la opinión del analista deberá usarse para determinar si el tamaño de muestra final debe ser mayor.

TAMAÑO DE MUESTRA PARA UNA ESTIMACIÓN POR INTERVALO DE LA MEDIA POBLACIONAL

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} \tag{8.3}$$

Este tamaño de muestra proporciona el margen de error deseado al nivel de confianza elegido.

En la ecuación (8.3) E es el margen de error que el usuario está dispuesto a aceptar, y el valor $z_{\alpha/2}$ es consecuencia directa del nivel de confianza que se va usar para calcular la estimación por intervalo. A reserva de la decisión del usuario, 95% de confianza es el valor más usado ($z_{0.025} = 1.96$).

Por último, para usar la ecuación (8.3) es necesario contar con el valor de la desviación estándar poblacional σ. Sin embargo, aun cuando este valor no se conozca, puede usarse la ecuación (8.3) siempre que se tenga un valor preliminar o un *valor planeado* de σ. En la práctica, se suele usar alguno de los procedimientos siguientes para obtener este valor planeado de σ.

El valor planeado de la desviación estándar poblacional σ debe especificarse antes de determinar el tamaño de la muestra. Aquí se muestran tres métodos para obtener este valor planeado de σ.

1. Usar como valor planeado de σ una estimación de la desviación estándar poblacional calculada a partir de datos de estudios anteriores.
2. Emplear un estudio piloto seleccionando una muestra preliminar. La desviación estándar muestral obtenida de la muestra preliminar puede usarse como valor planeado de σ.
3. Use su juicio para el valor de σ. Por ejemplo, se puede empezar por estimar el mayor y el menor valor en los datos de la población. Esta diferencia entre el mayor y el menor valor proporciona una estimación del rango de los datos. Por último, este valor dividido entre 4 suele considerarse como una aproximación burda a la desviación estándar y tomarse como un valor planeado aceptable de σ.

Para demostrar cómo se usa la ecuación (8.3) en la determinación del tamaño de la muestra, se considera el ejemplo siguiente. En un estudio previo para investigar el costo de la renta de automóviles en Estados Unidos se encontró que el costo medio de la renta de un automóvil mediano era aproximadamente $55 por día. Suponga que la organización que realizó dicho estudio quiere realizar un nuevo estudio para estimar la media poblacional de las rentas por día de automóviles medianos en Estados Unidos. Antes de iniciar, especificó que la media poblacional de las rentas por día debe estimarse con un margen de error de $2 y que se desea un nivel de 95% de confianza.

El margen de error especificado es $E = 2$, el nivel 95% de confianza indica que $z_{0.025} = 1.96$. Por tanto, sólo falta un valor planeado de la desviación estándar poblacional σ para calcular el tamaño de muestra deseado. El analista revisó los datos muestrales del estudio anterior y encon-

La ecuación (8.3) proporciona el tamaño de muestra mínimo necesario para obtener el margen de error deseado. Si el tamaño de muestra calculado no es un número entero, se redondea al siguiente número entero, con lo que se tendrá un margen de error ligeramente menor al requerido.

tró que la desviación estándar poblacional del costo de la renta diaria era \$9.65. Usando \$9.65 como valor planeado de σ, se tiene

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(1.96)^2(9.65)^2}{2^2} = 89.43$$

De esta manera, el tamaño de la muestra necesario para obtener un margen de error de \$2 debe ser de por lo menos 89.43 rentas de automóviles medianos. En casos como éste, en los que el valor de n no es un número entero, se redondea al siguiente valor entero; así que el tamaño de muestra que se aconseja es 90 rentas de automóviles medianos.

Ejercicios

Métodos

23. ¿Qué tan grande debe seleccionarse una muestra para tener un intervalo de confianza de 95% con un margen de error de 10? Suponga que la desviación estándar poblacional es 40.



24. En un conjunto de datos se estima que el rango es 36.
    a. ¿Cuál es el valor planeado para la desviación estándar poblacional?
    b. ¿De qué tamaño deberá ser la muestra para que el margen de error en un intervalo de confianza de 95% sea 3?
    c. ¿De qué tamaño deberá ser la muestra para que el margen de error en un intervalo de confianza de 95% sea 2?



Aplicaciones



25. Remítase al ejemplo de Scheer Industries de la sección 8.2. Como valor planeado para la desviación estándar poblacional use 6.84 día.
    a. Asuma 95% de confianza, ¿de qué tamaño deberá ser la muestra para tener un margen de error de 1.5 días?
    b. Si se desea un intervalo de 90% de confianza, ¿de qué tamaño deberá ser la muestra para tener un margen de error de 2 días?

26. El costo promedio de la gasolina sin plomo en Grater Cincinnati es \$2.41 (*The Cincinnati Enquirer*, 3 de febrero de 2006). En una época de cambios en los precios, un periódico muestrea las gasolineras y presenta un informe sobre los precios de la gasolina. Suponga que en los precios del galón de la gasolina sin plomo la desviación estándar es \$0.15; dé el tamaño de muestra n que debe usar este periódico para tener 95% de confianza con cada uno de los márgenes de error siguientes.
    a. Un margen de error de \$0.07
    b. Un margen de error de \$0.05
    c. Un margen de error de \$0.03

27. Los salarios anuales iniciales de estudiantes que acaban de terminar una carrera en administración se espera que estén entre \$30 000 y \$45 000. Suponga que quiere dar un intervalo de confianza de 95% para estimar la media poblacional de los salarios iniciales. ¿Cuál es el valor planeado de la desviación estándar poblacional? ¿Cuán grande deberá ser la muestra si quiere que el margen de error sea
    a. \$500?
    b. \$200?
    c. \$100?
    d. ¿Recomendaría usted tratar de tener \$100 como margen de error?

28. Smith Travel Research proporciona información sobre los precios por noche de las habitaciones de hotel en Estados Unidos (*USA Today*, 8 de julio de 2002). Use \$2 como el margen de error deseado y \$22.50 como valor planeado para la desviación estándar poblacional, y encuentre los tamaños de muestra que se solicitan en los incisos a, b y c.
    a. Para un intervalo de confianza de 90% estime el precio medio de las habitaciones de hotel.
    b. Para un intervalo de confianza de 95% estime el precio medio de las habitaciones de hotel.

c. Para un intervalo de confianza de 99% estime el precio medio de las habitaciones de hotel.
d. Cuando se tiene un margen de error fijo, ¿qué pasa con el tamaño de la muestra a medida que el nivel de confianza aumenta? ¿Le recomendaría a Smith Travel Research que use 99% como nivel de confianza? Discuta.

29. El *2003 Information Please Almanac* presenta los tiempos que para transportarse al trabajo son requeridos en las 15 ciudades más grandes de Estados Unidos. Suponga que usa una muestra aleatoria simple preliminar de los habitantes de San Francisco y como valor planeado para la desviación estándar poblacional obtiene 6.25 minutos.
   a. Si desea estimar la media poblacional del tiempo que necesitan en San Francisco para transportarse al trabajo, con un margen de error de 2 minutos, ¿cuál debe ser el tamaño de la muestra? Suponga que el nivel de confianza es de 95%.
   b. Si desea estimar la media poblacional del tiempo que se necesita en San Francisco para transportarse al trabajo, con un margen de error de 1 minuto, ¿cuál debe ser el tamaño de la muestra? Suponga que el nivel de confianza es de 95%.

30. El primer trimestre del 2003, la proporción precio/ganancia P/G en las acciones de la Bolsa de Nueva York iba de 5 a 60 (*The Wall Steer Journal*, 7 de marzo de 2003). Si se desea estimar la media poblacional de esta relación P/G en todas las acciones de la Bolsa de Nueva York, ¿cuántas acciones habrá que tomar en la muestra, si se quiere que el margen de error sea 3? Use 95% de confianza.



8.4 Proporción poblacional

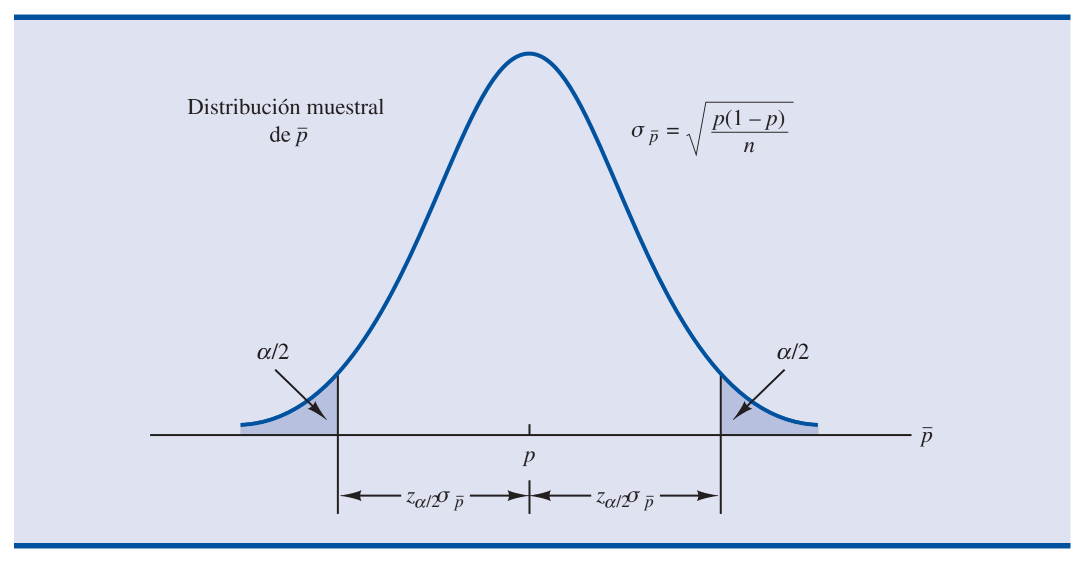
En la introducción a este capítulo se dijo que para obtener una estimación por intervalo de la proporción poblacional p, la fórmula general era

$$\bar{p} \pm \text{margen de error}$$

En el cálculo de esta estimación por intervalo la distribución muestral es importante.

En el capítulo 7 se dijo que la distribución muestral de $\bar{p}$ se aproxima mediante una distribución normal siempre que $np \geq 5$ y $n(1 - p) \geq 5$. En la figura 8.9 se muestra una aproximación

FIGURA 8.9 APROXIMACIÓN NORMAL A LA DISTRIBUCIÓN MUESTRAL DE $\bar{p}$



normal a la distribución muestral de $\bar{p}$. La media de la distribución muestral de $\bar{p}$ es la proporción poblacional p y el error estándar de $\bar{p}$ es

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{8.4}$$

Como la distribución muestral de $\bar{p}$ es una distribución normal, si en la estimación por intervalo de la proporción poblacional se elige como margen de error $z_{\alpha/2}\sigma_{\bar{p}}$, entonces $100(1-\alpha)\%$ de los intervalos que se obtengan contendrán la verdadera proporción poblacional. Pero para calcular el margen de error no se puede usar $\sigma_{\bar{p}}$, ya que no se conoce p; p es lo que se está tratando de estimar. Lo que se hace, es que p se sustituye por $\bar{p}$ y de esta manera se calcula el margen de error y la estimación por intervalo de la proporción poblacional queda dada por

$$\text{Margen de error} = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8.5}$$

Con este margen de error la expresión general para la estimación por intervalo de la proporción poblacional es la siguiente.

El margen de error para un intervalo de confianza para la proporción poblacional está dado por la cantidad $z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}$

ESTIMACIÓN POR INTERVALO DE UNA PROPORCIÓN POBLACIONAL

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8.6}$$

donde $1-\alpha$ es el coeficiente de confianza y $z_{\alpha/2}$ es el valor de z que deja un área $\alpha/2$ en la cola superior de la distribución normal estándar.



TeeTimes

En el siguiente ejemplo se ilustra el cálculo del margen de error y de la estimación por intervalo para una proporción muestral. Un estudio en Estados Unidos encuestó a 900 golfistas para conocer su opinión acerca de cómo se les trataba en los cursos de golf. En el estudio se encontró que 396 golfistas estaban satisfechas con la disponibilidad de horarios de salida. Por tanto, la estimación puntual de la proporción poblacional de golfistas satisfechas con la disponibilidad de horarios de salida es 396/900 = 0.44. Usando la expresión (8.6) y el nivel de 95% de confianza,

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$0.44 \pm 1.96\sqrt{\frac{0.44(1-0.44)}{900}}$$

$$0.44 \pm 0.0324$$

En consecuencia, el margen de error es 0.0324 y la estimación por intervalo de confianza de 95% de la proporción poblacional es 0.4076 a 0.4724. Empleando porcentajes, los resultados de la investigación permiten decir con 95% de confianza que entre 40.76 y 47.24% de las golfistas están satisfechas con la disponibilidad de horarios de salida.

Determinación del tamaño de la muestra

Ahora se considera cuál debe ser el tamaño de la muestra para obtener una estimación de la proporción poblacional con una precisión determinada. La función que tiene el tamaño de la muestra en la determinación de la estimación por intervalo de p es semejante a la que tiene en la estimación de la media poblacional estudiada en la sección 8.3. Ya en esta sección se dijo que el margen de error en la estimación por intervalo de la proporción poblacional es $z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}$. El margen de error se basa en el valor de $z_{\alpha/2}$, en la proporción muestral $\bar{p}$, y en el tamaño de la muestra n. Muestras mayores proporcionan márgenes de error menores y mejor precisión.

Sea E el margen de error deseado.

$$E = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Despejando n de esta fórmula se obtiene la fórmula para calcular el tamaño de la muestra con el que se tendrá el margen de error deseado, E.

$$n = \frac{(z_{\alpha/2})^2\bar{p}(1-\bar{p})}{E^2}$$

Sin embargo, debido a que no se conocerá $\bar{p}$ sino hasta que se tome la muestra, no es posible usar esta fórmula para calcular el tamaño de la muestra con el que se obtendrá el margen de error deseado. Lo que se necesita, entonces, es un valor planeado de $\bar{p}$ útil para hacer este cálculo. Con p^* como valor planeado de $\bar{p}$ la fórmula para calcular el tamaño de la muestra con el que se obtendrá el error E queda como se presenta a continuación.

TAMAÑO DE LA MUESTRA PARA UNA ESTIMACIÓN POR INTERVALO DE LA PROPORCIÓN POBLACIONAL

$$n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2} \tag{8.7}$$

En la práctica, el valor planeado p^* se determina mediante alguno de los métodos siguientes.

1. Utilizar la proporción poblacional de una muestra previa de las mismas unidades o de unidades similares.
2. Utilizar un estudio piloto y elegir una muestra preliminar. La proporción muestral de esta muestra se usa como valor planeado, p^*.
3. Proponer una "mejor aproximación" para el valor de p^*.
4. Si no aplica ninguna de las alternativas anteriores, emplear como valor planeado $p^* = 0.50$.

De regreso al estudio con mujeres golfistas, suponga que la empresa desea llevar a cabo otro estudio para determinar la proporción en la población de golfistas que está satisfecha con la disponibilidad de horarios de salida. ¿De qué tamaño deberá ser la muestra si se desea que en la estimación de la proporción poblacional el margen de error sea 0.025 a 95% de confianza? Como $E = .025$ y $z_{\alpha/2} = 1.96$, se necesita un valor planeado p^* para responder esta pregunta sobre el tamaño de la muestra. Utilizando como valor planeado p^*, el resultado del estudio anterior, $\bar{p} = 0.44$, con la ecuación (8.7) se obtiene

$$n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2} = \frac{(1.96)^2(0.44)(1-0.44)}{(0.025)^2} = 1514.5$$

TABLA 8.5 ALGUNOS VALORES DE $p^*(1 - p^*)$

p^*	$p^*(1 - p^*)$	
0.10	(0.10)(0.90) = 0.09	
0.30	(0.30)(0.70) = 0.21	
0.40	(0.40)(0.60) = 0.24	
0.50	(0.50)(0.50) = 0.25	← Mayor valor de $p^*(1 - p^*)$
0.60	(0.60)(0.40) = 0.24	
0.70	(0.70)(0.30) = 0.21	
0.90	(0.90)(0.10) = 0.09	

Así, el tamaño de la muestra debe ser por lo menos 1514.5 golfistas para tener el margen de error requerido. Redondeando al valor entero siguiente, se tiene que se necesitan 1515 golfistas para obtener el margen de error deseado.

La cuarta alternativa sugerida para seleccionar un valor planeado p^* *es* elegir $p^* = 0.50$. Cuando no se cuenta con ninguna otra información suele usarse este valor. Para entender la razón, observe que el numerador de la ecuación (8.7) indica que el tamaño de la muestra es proporcional a la cantidad $p^*(1 - p^*)$. Si el valor de $p^*(1 - p^*)$ es grande, el tamaño de la muestra será grande. En la tabla 8.5 se muestran algunos de los valores que puede tener $p^*(1 - p^*)$. El mayor valor de $p^*(1 - p^*)$ se presenta cuando $p^* = 0.50$. De esta manera, en caso de duda acerca del valor planeado apropiado, se sabe que $p^* = 0.50$ dará el mayor tamaño de muestra que se puede recomendar. En efecto, recomendando el mayor tamaño de muestra posible se va a lo seguro. Si resulta que la proporción muestral es diferente del valor planeado, el margen de error será menor que el deseado. De manera que al usar $p^* = 0.50$, se garantiza que el tamaño de la muestra será suficiente para obtener el margen de error deseado.

En el ejemplo del estudio de las golfistas, si se usa como valor planeado $p^* = 0.50$, el tamaño de muestra que se obtiene es

$$n = \frac{(z_{\alpha/2})^2 p^*(1 - p^*)}{E^2} = \frac{(1.96)^2(0.50)(1 - 0.50)}{(0.025)^2} = 1536.6$$

Es decir, una muestra ligeramente mayor, 1537 golfistas.

NOTAS Y COMENTARIOS

El margen de error deseado para calcular una proporción poblacional casi siempre es 0.10 o menos. En las encuestas de opinión pública a nivel nacional en Estados Unidos, conducidas por Gallup y Harris, un margen de error de 0.03 o 0.04 es común. Con dichos márgenes de error, la ecuación (8.7) suministra un tamaño de la muestra que es suficiente para satisfacer los requerimientos de $np \geq 5$ y $n(1 - p) \geq 5$ para usar una distribución normal como una aproximación de la distribución muestral de $\bar{x}$.

Ejercicios

Métodos



31. En una muestra aleatoria simple de 400 individuos, 100 de las respuestas fueron Sí.
    a. Dé la estimación puntual de la proporción poblacional de individuos cuya respuesta será Sí.
    b. Dé la estimación del error estándar de la proporción $\sigma_{\bar{p}}$.
    c. Calcule el intervalo de confianza de 95% para la proporción poblacional.

32. En una muestra aleatoria de 800 elementos se obtiene una proporción muestral, $\bar{p} = 0.70$.
    a. Dé un intervalo de 90% de confianza para la proporción poblacional.
    b. Encuentre un intervalo de confianza de 95% para la proporción poblacional.
33. En un estudio el valor planeado para la proporción poblacional es $p^* = 0.35$. ¿De qué tamaño se debe tomar la muestra para dar un intervalo de confianza de 95% con un margen de error de 0.05?
34. Para 95% de confianza, ¿de qué tamaño deberá tomar la muestra para obtener un margen de error de 0.03 en la estimación de una proporción poblacional? Suponga que no se cuenta con datos anteriores para obtener un valor planeado de p^*.

Aplicaciones



35. Se hizo un estudio con 611 oficinistas para investigar su atención al teléfono, el estudio registraba la frecuencia con que contestaban el teléfono y la frecuencia con que dejaban que la llamada pase al buzón de voz (*USA Today*, 21 de abril de 2002). De estos oficinistas, 281 indicaron constatar siempre las llamadas y no utilizar el buzón de voz.
    a. Dé la estimación puntual de la proporción poblacional de oficinistas que siempre responden el teléfono.
    b. A 90% de confianza, ¿cuál es el margen de error?
    c. Dé el intervalo de 90% de confianza para la proporción de la población de oficinistas que siempre contestan el teléfono.
36. De acuerdo con estadísticas publicadas por la CNBC, la cantidad de vehículos que no están asegurados es sorprendente (CNBC, 23 de febrero de 2006). Los resultados muestrales de la CNBC indican que 46 de 200 vehículos no estaban asegurados.
    a. ¿Cuál es la estimación puntual de la proporción de vehículos no asegurados?
    b. Dé un intervalo de confianza de 95% para la proporción poblacional.
37. Towers Perrin, una empresa de recursos humanos, realizó un estudio con 1100 empleados de empresas medianas y grandes para determinar qué tan insatisfechos estaban con sus trabajos (*The Wall Street Journal*, 29 de enero de 2003). En el archivo JobSatisfaction se muestran datos representativos. Un Sí como respuesta indica que al empleado le desagrada mucho su empleo actual.



    a. Dé la estimación puntual de la proporción poblacional de empleados a quienes les disgusta mucho su empleo actual.
    b. A 95% de confianza, ¿cuál es el margen de error?
    c. ¿Cuál es el intervalo de confianza de 95% para la proporción de la población de empleados a quienes les desagrada mucho su empleo actual?
    d. Towers Perrin estima que a los empleadores les cuesta un tercio de un sueldo anual por hora hallar un sucesor y hasta 1.5 veces el sueldo anual encontrar un sucesor para un empleado que recibe una compensación elevada. ¿Cuál es el mensaje de esta investigación para los empleadores?
38. Según Thomson Financial, hasta el 25 de enero de 2006, la mayor parte de las empresas que informaban tener ganancias habían superado las estimaciones (*BusinessWeek*, 6 de febrero de 2006). En una muestra de 162 empresas, 104 superaron las estimaciones, 29 coincidieron y 29 se quedaron cortas.
    a. ¿Cuál es la estimación puntual de la proporción de empresas que se quedaron cortas?
    b. Determine el margen de error y dé un intervalo de confianza de 95% para la proporción que superó las estimaciones.
    c. ¿De qué tamaño debe de ser la muestra si el margen de error es 0.05?



39. En 2003 el porcentaje de personas que no tenía un seguro médico (en Estados Unidos) era 15.6% (*Statistical Abstract of the United Status*, 2006). Se le pide a un comité del Congreso realizar un estudio para obtener información actualizada.
    a. ¿Qué tamaño de muestra le recomienda usted al comité, si el objetivo es que en la estimación de la proporción actual de individuos que no tienen seguro médico el margen de error sea 0.03? Use 95% de confianza.
    b. Repita el inciso a usando 99% de confianza.

40. En el béisbol profesional el récord de cuadrangulares era de 61 conectados en una temporada y perteneció durante 37 años a Roger Maris de los Yankees de Nueva York. Sin embargo, de 1998 a 2001, tres jugadores —Mark McGwire, Sammy Sosa y Barry Bonds— rompieron este récord, siendo Bonds quien tiene el récord actual de 73 cuadrangulares en una sola temporada. Debido al rompimiento de este récord después de tanto tiempo, y también al rompimiento de otros récords, se sospecha que los jugadores de béisbol emplean medicamentos ilegales conocidos como esteroides para la formación de músculo. En una encuesta realizada por *USA Today*/CNN/Gallup se encontró que 86% de los aficionados al béisbol opinaba que a los jugadores profesionales se les debería someter a un examen para detectar esteroides (*USA Today*, 18 de julio de 2002). Si en un estudio se seleccionan 650 aficionados al béisbol, calcule el margen de error y el intervalo de confianza de 95% para la proporción poblacional de aficionados que opinan que los jugadores de béisbol profesional deberían ser sometidos a un examen para detectar esteroides.
41. La juventud de Estados Unidos usa Internet intensamente; el 87% de los jóvenes entre 12 y 17 años son usuarios de Internet (*The Cincinnati Enquirer*, 7 de febrero de 2006). En una muestra de usuarios de Internet de esta edad, 9% votó por MySpace como el sitio de Internet más popular. Suponga que en este estudio participaron 1 400 jóvenes. ¿Cuáles son los márgenes de error y la estimación por intervalo de la proporción poblacional de quienes consideran que este sitio es el más popular? Use 95% de confianza.
42. Una encuesta realizada por *USA Today*/CNN/Gallup durante la campaña presidencial tomó en junio una muestra de 491 votantes potenciales (*USA Today*, 9 de junio de 2000). El objetivo de esta encuesta era estimar la proporción de votantes potenciales a favor de cada candidato. Suponga que el valor planeado es $p^* = 0.50$ con un nivel de confianza de 95%.
    a. Si $p^* = 0.50$. ¿Cuál fue el margen de error planeado en la encuesta de junio?
    b. Al acercarse la elección de noviembre se busca una mejor precisión y un menor margen de error. Suponga que los márgenes de error que se piden son los que se muestran en la tabla siguiente. Calcule el tamaño de muestra que se recomienda para cada estudio.

Estudio	**Margen de error**
Septiembre	0.04
Octubre	0.03
Comienzo de noviembre	0.02
Un día antes de la elección	0.01

43. Phoenix Wealth Management/Harris realizó un estudio con 1500 individuos cuyo patrimonio era de un millón o más de dólares, obtuvo diversos estadísticos sobre la gente rica (*BusinessWeek*, 22 de septiembre de 2003). Los tres años anteriores habían sido malos para el mercado de acciones, lo que motivó algunas de las preguntas realizadas.
    a. En este estudio se encontró que 53% de los encuestados perdió 25% o más del valor de su portafolio en los últimos tres años. Dé un intervalo de confianza de 95% para la proporción de gente rica que perdió 25% o más del valor de su portafolio en los últimos tres años.
    b. El estudio indicó que 31% de los encuestados siente que deberá ahorrar más para su retiro para compensar lo perdido. Dé un intervalo de confianza de 95% para la proporción poblacional.
    c. Cinco por ciento de los encuestados hicieron una donación de $25 000 o más para obras de caridad el año anterior. Dé un intervalo de confianza de 95% para la proporción de quienes hicieron una donación de $25 000 o más para obras de caridad.
    d. Compare los márgenes de error de las estimaciones por intervalo de los incisos a, b y c. ¿Cuál es la relación entre margen de error y $\bar{p}$? Si usa la misma muestra para obtener varias proporciones, ¿cuál de las proporciones debe usarse para estimar el valor planeado p^*? ¿Por qué considera que en estos casos suela usarse $p^* = 0.50$?

Resumen

En este capítulo se presentaron los métodos para obtener estimaciones por intervalo de la media poblacional y de la proporción poblacional. Un estimador puntual puede o no proporcionar una buena estimación de un parámetro poblacional. Un intervalo de estimación suministra una medida de

la precisión de una estimación. Tanto la estimación por intervalo de una media poblacional como la de una proporción poblacional tienen la forma: estimación puntual ± margen de error.

Para la media poblacional se presentaron estimaciones por intervalo en dos casos. En el caso σ conocida, se usan datos históricos o alguna otra información para obtener una estimación de σ antes de tomar la muestra. Entonces, el análisis de nuevos datos muestrales se realiza bajo la suposición de que se conoce σ. En el caso σ desconocida, los datos muestrales se usan para estimar tanto la media poblacional como la desviación estándar poblacional. La decisión final de qué procedimiento de estimación por intervalo usar depende de que el analista decida qué método proporciona una mejor estimación de σ.

En el caso σ conocida, el procedimiento de estimación por intervalo se basa en el valor supuesto para σ y en el uso de la distribución normal estándar. En el caso σ desconocida, para el procedimiento de estimación por intervalo se usa la desviación estándar muestral s y la distribución t. En ambos casos, la calidad de la estimación por intervalo depende de la distribución de la población y del tamaño de la muestra. Si la población tiene una distribución normal, la estimación por intervalo será exacta en ambos casos, aun cuando los tamaños de las muestras sean pequeños. Si la población no tiene una distribución normal, la estimación por intervalo resultante será aproximada. Tamaños de muestras mayores proporcionarán mejores aproximaciones, pero entre más sesgada sea la población, mayor será el tamaño de la muestra necesario para obtener una buena aproximación. En las secciones 8.1 y 8.2 se dieron consejos prácticos respecto al tamaño de muestra necesario para obtener buenas aproximaciones. En la mayoría de los casos, una muestra de tamaño 30 o mayor proporcionará una buena aproximación para el intervalo de confianza.

La forma general de una estimación por intervalo para la proporción poblacional es $\bar{p}$ ± margen de error. En la práctica, los tamaños de muestra empleados en estimaciones por intervalo de una proporción poblacional suelen ser grandes. Entonces, el procedimiento de estimación por intervalo se basa en la distribución normal estándar.

Algunas veces se suele especificar un determinado margen de error antes de llevar a cabo el plan de muestreo. También se explicó cómo elegir el tamaño de muestra adecuado para obtener la precisión deseada.

Glosario

Estimación por intervalo Estimación de un parámetro poblacional que suministra un intervalo que se cree contiene el valor del parámetro. Para las estimaciones por intervalo vistas en este capítulo tiene la forma: estimación puntual ± margen de error.

Margen de error Valor que se resta y se suma a la estimación puntual con objeto de obtener un intervalo de estimación para el parámetro poblacional.

σ conocida Caso en el que datos históricos u alguna otra información proporciona un buen valor para ser considerado como desviación estándar poblacional antes de tomar la muestra. Este valor conocido de σ se usa en la estimación por intervalo para calcular el margen de error.

Nivel de confianza Confianza correspondiente a la estimación por intervalo. Por ejemplo, si un procedimiento para obtener una estimación por intervalo proporciona intervalos tales que, 95% de ellos contendrán al parámetro poblacional, se dice que esa estimación por intervalo tiene un nivel de confianza de 95%.

Coeficiente de confianza El nivel de confianza expresado como valor decimal. Por ejemplo 0.95 es el coeficiente de confianza correspondiente al nivel de confianza de 95%.

Intervalo de confianza Otro nombre para una estimación por intervalo.

σ desconocida El caso más común cuando no existen bases sólidas para estimar la desviación estándar poblacional antes de tomar la muestra. En la estimación por intervalo se usa la desviación estándar muestral para calcular el margen de error.

Distribución t Una familia de distribuciones de probabilidad que se usa para obtener una estimación por intervalo de la media poblacional cuando la desviación estándar poblacional σ no se conoce y se estima mediante la desviación estándar muestral s.

Grados de libertad Parámetro de las distribuciones t. Cuando se usa una distribución t para calcular una estimación por intervalo de la media poblacional, la distribución t correspondiente tiene $n - 1$ grados de libertad, donde n es el tamaño de la muestra aleatoria simple.

Fórmulas clave

Estimación por intervalo de una media poblacional: σ conocida

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{8.1}$$

Estimación por intervalo de una media poblacional: σ desconocida

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \tag{8.2}$$

Tamaño de la muestra para una estimación de la media poblacional

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2} \tag{8.3}$$

Estimación por intervalo de una proporción poblacional

$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}} \tag{8.6}$$

Tamaño de la muestra para una estimación de la proporción poblacional

$$n = \frac{(z_{\alpha/2})^2 p^*(1 - p^*)}{E^2} \tag{8.7}$$

Ejercicios complementarios

44. En un estudio realizado con 54 corredores de bolsa con descuento, se encontró que la media de los precios cobrados por una transacción de 100 acciones a $50 la acción, fue $33.77 (*AAII Journal*, febrero de 2006). Este estudio se realiza anualmente. De acuerdo con los datos históricos disponibles, considere que la desviación estándar poblacional conocida es $15.
    a. Según los datos muestrales, ¿cuál es el margen de error en un intervalo de confianza de 95%?
    b. Dé un intervalo de confianza de 95% para la media de los precios cobrados por una transacción de 100 acciones a $50 la acción.
45. En un estudio realizado por la American Automobile Association se encontró que una familia de cuatro miembros en vacaciones gasta en promedio $215.60 por día. Suponga que en una muestra de 64 familias de vacaciones en las Cataratas del Niágara la media muestral encontrada haya sido $252.45 por día y la desviación estándar muestral $74.50.
    a. Dé una estimación, mediante un intervalo de confianza de 95% para la media de la cantidad que gasta por día una familia de cuatro, que está de vacaciones en las Cataratas del Niágara.
    b. De acuerdo con el intervalo de confianza del inciso a, ¿parece que la media poblacional de la cantidad gastada por día por las familias que visitan las Cataratas del Niágara es diferente de la media reportada por la American Automobile Association? Explique.
46. La película *Harry Poter y la piedra filosofal* echó por tierra el récord de taquilla en estreno de *El mundo perdido: Parque Jurásico* (*The WallStreet Journal*, 19 de noviembre de 2001). En una muestra de 100 cines se encontró que la media de la recaudación bruta en los tres días del primer fin de semana fue de $25 467 por cine. La desviación estándar fue $4980.
    a. ¿Cuál es el margen de error en este estudio? Use 95% de confianza.
    b. ¿Cuál es la estimación del intervalo de confianza de 95% de la media poblacional de la recaudación por cine?
    c. *El mundo perdido* obtuvo $72.1 millones en los tres días del primer fin de semana. *Harry Poter y la piedra filosofal* se presentó en 3672 cines. Dé una estimación del total obtenido por *Harry Poter y la piedra filosofal* en los tres días del primer fin de semana.
    d. En un artículo de Associated Press se dijo que *Harry Poter* "echó por tierra" el récord de taquilla en el debut de *El mundo perdido*. ¿Coinciden sus resultados con esto?

47. Muchos observadores de los mercados de acciones aseguran que cuando la proporción P/E en las acciones es superior a 20, el mercado está sobrevaluado. La proporción P/E es el precio de una acción dividido entre las ganancias de los últimos 12 meses. Suponga que usted desea saber si actualmente el mercado está sobrevaluado y qué proporción de las empresas pagan dividendos. A continuación aparece una lista de 30 empresas que cotizan en la Bolsa de Nueva York (NYSE) (*Barron's*, 19 de enero de 2004).

archivo en CD
NYSEStocks

Empresa	Dividendos	Proporción P/E	Empresa	Dividendos	Proporción P/E
Albertsons	Sí	14	NY Times A	Sí	25
BRE Prop	Sí	18	Omnicare	Sí	25
CityNtl	Sí	16	PallCp	Sí	23
DelMonte	No	21	PubSvcEnt	Sí	11
EnrgzHldg	No	20	SensientTch	Sí	11
Ford Motor	Sí	22	SmtProp	Sí	12
Gildan A	No	12	TJX Cos	Sí	21
HudsnUtdBcp	Sí	13	Thomson	Sí	30
IBM	Sí	22	USB Hldg	Sí	12
JeffPilot	Sí	16	US Restr	Sí	26
KingswayFin	No	6	Varian Med	No	41
Libbey	Sí	13	Visx	No	72
MasoniteIntl	No	15	Waste Mgt	No	23
Motorola	Sí	68	Wiley A	Sí	21
Ntl City	Sí	10	Yum Brands	No	18

a. Dé una estimación puntual para la proporción poblacional P/E en las acciones que cotizan en la Bolsa de Nueva York. Dé un intervalo de confianza de 95%.
b. De acuerdo con su respuesta al inciso a, ¿considera usted que el mercado está sobrevaluado?
c. Dé una estimación puntual de la proporción de empresas en la NYSE que pagan dividendos. ¿El tamaño de la muestra es suficientemente grande para justificar el empleo de la distribución normal en el cálculo de un intervalo de confianza para esta proporción? ¿Por qué sí o por qué no?




48. US Airways llevó a cabo diversos estudios que indican ahorros importantes si los viajeros frecuentes del programa Dividend Miles realizaran en línea el canje de millas y programaran los vuelos ganados (*US Airways Attaché*, febrero de 2003). En un estudio se recogieron datos sobre el tiempo que se requiere para realizar por teléfono el canje de millas y la programación de un vuelo ganado. En el conjunto de datos Flights se encuentra una muestra de tiempos en minutos requeridos para programar por teléfono cada uno de 150 vuelos ganados. Use Minitab o Excel para contestar las preguntas siguientes.
a. ¿Cuál es la media muestral del número de minutos que se requiere para programar por teléfono los vuelos ganados?
b. Dé el intervalo de confianza de 95% para la media poblacional del tiempo requerido para programar por teléfono los vuelos.
c. Suponga que un agente de boletos por teléfono trabaja 7.5 horas. ¿Cuántos vuelos ganados se espera que atienda dicho agente en un día?
d. Diga cómo apoya esta información al plan de US Airways de usar un sistema en línea para reducir costos.



49. En un estudio se les pidió a 200 ejecutivos de una muestra proporcionar datos sobre la cantidad de minutos por día que pierden los oficinistas tratando de localizar cosas mal guardadas, mal archivadas o mal clasificadas. Los datos de esta investigación se encuentran en el conjunto de datos ActTemps.
a. Use ActTemps para dar una estimación puntual de los minutos por día perdidos por los oficinistas en localizar cosas mal guardadas, mal archivadas o mal clasificadas.
b. ¿Cuál es la desviación estándar muestral?
c. Dé un intervalo de confianza de 95% para la cantidad de minutos perdidos por día.

50. Se hacen pruebas de rendimiento de gasolina con un determinado modelo de automóvil. Si se quiere dar un intervalo de confianza de 98% con un margen de error de 1 milla por galón, ¿cuán-

tos automóviles deberán usarse? Suponga que por pruebas anteriores se sabe que la desviación estándar del rendimiento es 2.6 millas por galón.

51. Un centro médico quiere estimar la media del tiempo que se necesita para programar una cita de un paciente. ¿De qué tamaño deberá ser la muestra si se quiere que el margen de error sea de dos minutos y que el nivel de confianza sea 95%? ¿De qué tamaño deberá tomarse la muestra si se quiere que el nivel de confianza sea 99%? Para la desviación estándar poblacional use como valor planeado, 8 minutos.

52. *BusinessWeek* presenta datos sobre el salario anual más bonos de presidentes de consejos de administración. En una muestra preliminar la desviación estándar es $675; los datos se dan en miles de dólares. ¿De cuántos presidentes de consejos de administración deberá constar la muestra si se quiere estimar el salario anual más bonos con un margen de error de $100 000? (*Nota:* El margen de error deseado será $E = 100$ si los datos están dados en miles de dólares.) Use 95% de confianza.

53. El National Center for Education Statistics (informa que 47% de los estudiantes universitarios trabaja para pagar sus estudios y su sustento. Suponga que se empleó una muestra de 450 estudiantes universitarios en ese estudio.
    a. Dé un intervalo de confianza de 95% para la proporción poblacional de estudiantes que trabajan para mantenerse y pagar sus estudios.
    b. Dé un intervalo de confianza de 99% para la proporción poblacional de estudiantes que trabajan para mantenerse y pagar sus estudios.
    c. ¿Qué ocurre con el margen de error cuando el nivel de confianza aumenta de 95% a 99%?

54. En un estudio de *USA Today*/CNN/Gallup, realizado con 369 padres que trabajan, se encontró que 200 consideran que pasan muy poco tiempo con sus hijos debido al trabajo.
    a. Dé una estimación puntual de la proporción poblacional de padres que trabajan y piensan que pasan muy poco tiempo con sus hijos debido al trabajo.
    b. ¿Cuál es el margen de error para 95% de confianza?
    c. ¿Cuál es el intervalo de confianza de 95% para la proporción poblacional de padres que trabajan y piensan que pasan muy poco tiempo con sus hijos debido al trabajo?

55. ¿De qué le sería más difícil deshacerse: de su televisión o de su computadora? En un estudio reciente efectuado con 1677 usuarios de Internet en Estados Unidos, se encontró que a 74% de la elite juvenil (edad promedio de 22 años) le sería más difícil deshacerse de su computadora (*PC Magazine*, 3 de febrero de 2004). Sólo para el 48% sería más difícil deshacerse de su televisión.
    a. Dé un intervalo de confianza de 95% para la proporción de jóvenes a quienes les sería más difícil deshacerse de su computadora.
    b. Encuentre un intervalo de confianza de 99% para la proporción de jóvenes a quienes les sería más difícil deshacerse de su televisión.
    c. ¿En cuál de los incisos a o b es mayor el margen de error? Explique por qué.

56. El aeropuerto internacional Cincinnati/Northern Kentucky tuvo en 2005 el segundo lugar en puntualidad en la llegada de vuelos entre los aeropuertos con más actividad del país (*The Cincinnati Enquirer*, 3 de febrero de 2003). Suponga que esto se basa en una muestra de 550 vuelos de los cuales 455 llegaron a tiempo.
    a. Elabore una estimación puntual de la tasa de llegadas a tiempo (proporción de vuelos que llegan a tiempo) al aeropuerto.
    b. Construya un intervalo de confianza de 95% para la proporción de llegadas a tiempo en la población de todos los vuelos del aeropuerto en 2005.

57. El *2003 Statistical Abstract of the United States* da el porcentaje de personas de 18 años o más que fuma. Asuma que en un nuevo estudio para recoger datos sobre los fumadores y no fumadores se usa como estimación preliminar de la proporción que fuma, 0.30.
    a. ¿De qué tamaño deberá tomarse la muestra para estimar la proporción de fumadores con un margen de error de 0.02? Use 95% de confianza.
    b. Suponga que el estudio usa su recomendación para el tamaño de la muestra del inciso a y encuentra 520 fumadores. ¿Cuál es la estimación puntual de la proporción de fumadores en la población?
    c. ¿Cuál es el intervalo de confianza de 99% para la proporción de fumadores en la población?

58. Una firma de tarjetas de crédito de un banco conocido desea estimar la proporción de tarjetahabientes que al final del mes tienen un saldo distinto de cero que ocasiona cargos. Suponga que el margen de error deseado es 0.03 con 98% de confianza.
    a. ¿De qué tamaño deberá tomarse la muestra si se cree que 70% de los tarjetahabientes de la firma tienen un saldo distinto de cero al final del mes?
    b. ¿De qué tamaño deberá tomarse la muestra si no se puede dar ningún valor planeado para la proporción?
59. En un estudio se le pidió a 200 personas que indicaran su mayor fuente de información de noticias; 110 indicaron que su principal fuente de noticias eran los noticieros de televisión.
    a. Dé un intervalo de confianza de 95% para la proporción poblacional de personas que tienen como principal fuente de noticias a la televisión.
    b. ¿Cuál será el tamaño de muestra necesario para estimar la proporción poblacional con un margen de error de 0.05 y 95% de confianza?
60. Aunque para los viajeros de negocios, los horarios y los costos son factores importantes al elegir una línea aérea, en un estudio realizado por *USA Today* se encontró que para los viajeros de negocios el factor más importante es que la línea tenga un programa de viajero frecuente. En una muestra de $n = 1993$ pasajeros que participaron en el estudio, 618 indicaron como factor más importante un programa de pasajero frecuente.
    a. ¿Cuál es la estimación puntual de la proporción poblacional de viajeros de negocios que consideran al programa de viajero frecuente como el factor más importante al elegir una línea aérea?
    b. Dé un intervalo de confianza de 95% para estimar la proporción poblacional.
    c. ¿De qué tamaño deberá ser la muestra para que el margen de error sea 0.01 con 95% de confianza? ¿Aconsejaría que *USA Today* tratara de tener esta precisión? ¿Por qué sí o por qué no?

Caso problema 1 La revista *Young Professional*

La revista *Young Professional* fue creada para un público formado por personas que se encuentran en los 10 primeros años de su carrera profesional en negocios. En sus dos primeros años de publicación, la revista ha tenido bastante éxito. Ahora el editor está tratando de aumentar la base publicitaria de su revista. Los anunciantes potenciales preguntan continuamente sobre los datos demográficos e intereses de los suscriptores de *Young Professional*. Para recabar esta información, la revista realizó un estudio para elaborar el perfil de sus suscriptores. Los resultados de dicho estudio se usarán para ayudar a elegir artículos de interés y para proporcionar a los anunciantes un perfil de los suscriptores. Como nuevo empleado de la empresa se le pide a usted su ayuda para analizar los resultados del estudio.

A continuación se presentan algunas de las preguntas del estudio.



1. Edad___
2. Sexo____
3. ¿Piensa comprar algún bien inmueble en los próximos dos años? Sí__ No__
4. ¿Cuál es el valor aproximado de las inversiones financieras, excluyendo su casa, que son propiedad suya o de otro miembro de su hogar?
5. ¿Cuántas transacciones de acciones/bonos/fondos mutualistas realizó el año pasado?
6. ¿Tiene en casa acceso de banda ancha a Internet? Sí ___ No ___
7. Por favor indique cuál fue el ingreso de su hogar el año pasado.
8. ¿Tiene hijos? Sí__ No___

El archivo Professional contiene las respuestas a estas preguntas. En la tabla 8.6 se muestra la parte de este archivo correspondiente a las respuestas de los primeros cinco entrevistados. El archivo completo se encuentra en el disco compacto que se distribuye con el libro.

TABLA 8.6 RESULTADOS PARCIALES DEL ESTUDIO DE LA REVISTA *YOUNG PROFESSIONAL*

Edad	Género	Compra de inmuebles	Valor de inversiones ($)	Número de transacciones	Conexión de banda ancha	Ingreso del hogar ($)	Hijos
38	Femenino	No	12 200	4	Sí	75 200	Sí
30	Masculino	No	12 400	4	Sí	70 300	Sí
41	Femenino	No	26 800	5	Sí	48 200	No
28	Femenino	Sí	19 600	6	No	95 300	No
31	Femenino	Sí	15 100	5	No	73 300	Sí
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Informe administrativo

Elabore un informe administrativo dando los resultados del estudio. Además de los resúmenes estadísticos, analice cómo la revista puede usar estos resultados para atraer más anunciantes. También presente una recomendación a los editores para que empleen los resultados del estudio en la elección de los temas de interés para sus suscriptores. Su informe debe contener los siguientes puntos, pero no limite su análisis a estas áreas.

1. Dé los estadísticos descriptivos adecuados para resumir los datos.
2. Muestre los intervalos de 95% de confianza para la edad promedio y para el ingreso promedio por hogar de los suscriptores.
3. Encuentre intervalos de confianza de 95% para la proporción de suscriptores que tienen acceso de banda ancha y para la proporción de éstos que tienen niños.
4. ¿Será *Young Professional* un buen sitio para la publicidad de agentes de bolsa? Justifique su conclusión con datos estadísticos.
5. ¿Será esta revista un buen lugar para la publicidad de empresas que venden software educativo y juegos de computadora para niños?
6. Haga un comentario sobre el tipo de artículos que crea usted son de interés para los lectores de la revista.

Caso problema 2 Gulf Real Estate Properties

Gulf Real Estate Properties, Inc., es una empresa inmobiliaria ubicada en el suroeste de Florida. Esta empresa, que se anuncia como "experta en el mercado de bienes raíces", monitorea las ventas de condominios recabando datos sobre ubicación, precio de lista, precio de venta y días necesarios para vender cada unidad. Los condominios están calificados como *con vista al golfo* o *sin vista al golfo* dependiendo de su ubicación hacia el Golfo de México. Multiple Listing Service Naples, Florida, proporciona datos muestrales sobre 40 condominios con vista al golfo y 18 condominios sin vista al golfo.* Los precios están dados en miles de dólares. Estos datos se presentan en la tabla 8.7.

Informe administrativo

1. Use los estadísticos descriptivos adecuados para resumir cada una de las tres variables de los 40 condominios con vista al golfo.
2. Aplique los estadísticos descriptivos adecuados para resumir cada una de las tres variables de los 18 condominios sin vista al golfo.
3. Compare los resultados. Analice cualquier estadístico específico que ayude al agente de ventas inmobiliarias a saber más sobre el mercado de los condominios.

*Datos sustentados en las ventas de condominios reportadas en el Naples MLS (Coldwell Banker, junio de 2000).

TABLA 8.7 DATOS DE VENTA DE PROPIEDADES VENDIDAS POR GULF REAL ESTATE PROPERTIES



Condominios con vista al golfo			Condominios sin vista al golfo		
Precio de lista	Precio de venta	Días hasta la venta	Precio de lista	Precio de venta	Días hasta la venta
495.0	475.0	130	217.0	217.0	182
379.0	350.0	71	148.0	135.5	338
529.0	519.0	85	186.5	179.0	122
552.5	534.5	95	239.0	230.0	150
334.9	334.9	119	279.0	267.5	169
550.0	505.0	92	215.0	214.0	58
169.9	165.0	197	279.0	259.0	110
210.0	210.0	56	179.9	176.5	130
975.0	945.0	73	149.9	144.9	149
314.0	314.0	126	235.0	230.0	114
315.0	305.0	88	199.8	192.0	120
885.0	800.0	282	210.0	195.0	61
975.0	975.0	100	226.0	212.0	146
469.0	445.0	56	149.9	146.5	137
329.0	305.0	49	160.0	160.0	281
365.0	330.0	48	322.0	292.5	63
332.0	312.0	88	187.5	179.0	48
520.0	495.0	161	247.0	227.0	52
425.0	405.0	149			
675.0	669.0	142			
409.0	400.0	28			
649.0	649.0	29			
319.0	305.0	140			
425.0	410.0	85			
359.0	340.0	107			
469.0	449.0	72			
895.0	875.0	129			
439.0	430.0	160			
435.0	400.0	206			
235.0	227.0	91			
638.0	618.0	100			
629.0	600.0	97			
329.0	309.0	114			
595.0	555.0	45			
339.0	315.0	150			
215.0	200.0	48			
395.0	375.0	135			
449.0	425.0	53			
499.0	465.0	86			
439.0	428.5	158			

4. Dé un intervalo de confianza de 95% para estimar las medias poblacionales del precio de venta y del número de días necesarios para vender los condominios con vista al golfo. Interprete los resultados.
5. Encuentre un intervalo de confianza de 95% para estimar las medias poblacionales del precio de venta y el número de días necesarios para vender los condominios sin vista al golfo. Interprete los resultados.
6. Suponga que se necesita estimar el precio medio de venta de los condominios con vista al golfo con un margen de error de $40 000 y la media del precio de venta de los condo-

minios sin vista al golfo con un margen de error de $15 000. Si se usa 99% de confianza, ¿de qué tamaño deberán ser las muestras?

7. La inmobiliaria Golfo Real firmó contratos para dos nuevos catálogos: un condominio con vista al Golfo con un precio de lista de $585 000 y un condominio sin vista al Golfo con un precio de $285 000. ¿Cuál es su estimado del precio final de venta y el número de días requerido para vender cada una de estas unidades?

Caso problema 3 Metropolitan Research, Inc.

Metropolitan Research, Inc., una organización para la investigación del consumo, realiza estudios que tienen por objeto evaluar una amplia variedad de productos y servicios para los consumidores. En uno de sus estudios, Metropolitan se enfocó en la satisfacción de los consumidores con el funcionamiento de los automóviles producidos por el principal fabricante de Detroit. En un cuestionario enviado a propietarios de automóviles de esta empresa se encontraron varias quejas relacionadas con problemas tempranos en la transmisión. Para tener más información acerca de los problemas en la transmisión, Metropolitan empleó una muestra de reparaciones de la transmisión proporcionada por empresas en Detroit, dedicadas a la reparación de transmisiones. Los datos siguientes son el número de milla recorridos por 50 automóviles hasta el momento en que se presentaron los problemas con la transmisión.



Auto

85 092	32 609	59 465	77 437	32 534	64 090	32 464	59 902
39 323	89 641	94 219	116 803	92 857	63 436	65 605	85 861
64 342	61 978	67 998	59 817	101 769	95 774	121 352	69 568
74 276	66 998	40 001	72 069	25 066	77 098	69 922	35 662
74 425	67 202	118 444	53 500	79 294	64 544	86 813	116 269
37 831	89 341	73 341	85 288	138 114	53 402	85 586	82 256
77 539	88 798						

Informe administrativo

1. Use los estadísticos descriptivos adecuados para resumir los datos sobre los problemas en la transmisión.
2. Dé un intervalo de confianza de 95% para estimar el número de millas promedio, en la población de automóviles con fallas en la transmisión, recorridas hasta que se presenta el problema. Haga una interpretación administrativa del intervalo estimado.
3. Analice las consecuencias de sus hallazgos en términos de la creencia de que algunos propietarios de automóviles tuvieron problemas tempranos con la transmisión.
4. ¿Cuántos registros deben tomarse en la muestra si se desea estimar la media poblacional del número de millas recorridas hasta la aparición de problemas en la transmisión con un margen de error de 5000 millas? Use 95% de confianza.
5. ¿Qué otra información desearía recolectar para evaluar mejor los problemas con la transmisión?

Apéndice 8.1 Estimación por intervalo con Minitab

A continuación se describe cómo usar Minitab para obtener intervalos de confianza para la media poblacional y para la proporción poblacional.

Media poblacional: σ conocida



Lloyd's

La estimación por intervalo se ilustra mediante el ejemplo de Lloyd's de la sección 8.1. En una muestra de 100 clientes las cantidades gastadas en cada visita a la tienda están en la columna C1 de la hoja de cálculo de Minitab. Se supone que la desviación estándar poblacional se conoce y que es $\sigma = 20$. Los pasos siguientes se usan para calcular un intervalo de confianza de 95% para estimar la media poblacional. Los datos están en la columna C1 de la hoja de cálculo de Minitab.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Basic Statistics**
Paso 3. Elegir **1-Sample Z**
Paso 4. Cuando aparezca el cuadro de diálogo 1-Sample Z:
Ingresar C1 en el cuadro **Samples in column**
Ingresar 20 en el cuadro **Standard deviation**
Paso 5. Clic en **OK**

Si no se especifica otra cosa, Minitab emplea 95% como nivel de confianza. Para especificar otro nivel de confianza, por ejemplo 90%, al paso 4 hay que agregar lo siguiente.

Seleccionar **Options**
Cuando aparezca el cuadro de diálogo 1-Sample Z:
Ingresar 90 en el cuadro **Confidence level**
Clic en **OK**

Media poblacional: σ desconocida



NewBalance

La estimación por intervalo se ilustra empleando los datos de la tabla 8.2 que dan los saldos en las tarjetas de crédito en una muestra de 70 hogares. Los datos están en la columna C1 de la hoja de cálculo de Minitab. En este caso se estima la desviación estándar poblacional σ mediante la desviación estándar muestral s. Con los pasos siguientes se obtiene un intervalo de confianza de 95% para estimar la media poblacional.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Basic Statistics**
Paso 3. Elegir **1-Sample t**
Paso 4. Cuando aparezca el cuadro de diálogo 1-Sample t
Ingresar C1 en el cuadro **Samples in columns**
Paso 5. Clic en **OK**

Si no se especifica otra cosa, Minitab emplea 95% como nivel de confianza. Para especificar otro nivel de confianza, por ejemplo 90% al paso 4 hay que agregar lo siguiente.

Seleccionar **Options**
Cuando aparezca el cuadro de diálogo 1-Sample t-Options:
Ingresar 90 en el cuadro **Confidence level**
Clic en **OK**

Proporción poblacional



TeeTimes

La estimación por intervalo se ilustra empleando los datos de las golfistas, presentados en la sección 8.4. Los datos están en la columna C1 de la hoja de cálculo de Minitab. Las respuestas se registraron como Sí si la golfista está satisfecha con la disponibilidad de horarios de salida y No, si no es el caso. Usando los pasos siguientes se calcula un intervalo de confianza de 95% para estimar la proporción de golfistas satisfechas con la disponibilidad de los horarios de salida.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Basic Statistics**
Paso 3. Elegir **1 Proportion**
Paso 4. Cuando aparezca el cuadro de diálogo 1 Proportion:
Ingresar C1 en el cuadro **Samples in columns**
Paso 5. Seleccionar **Options**
Paso 6. Cuando aparezca el cuadro de diálogo 1 Proportion-Options:
Seleccionar **Use test and interval based on normal distribution**
Clic en **OK**
Paso 7. Clic en **OK**

Si no se especifica otra cosa, Minitab emplea 95% como nivel de confianza. Para especificar otro nivel de confianza, como 90%, cuando aparezca el cuadro de diálogo 1 Proportions-Options en el paso 6, ingresar 90 en el cuadro **Confidence Level**.

Nota: La rutina 1 Proporción de Minitab usa de las respuestas puestas en orden alfabético y selecciona la *segunda respuesta* como la proporción poblacional de interés. En el ejemplo de las golfistas, Minitab usa el orden alfabético No-Sí y de esta manera da el intervalo de confianza para la proporción de respuestas Sí. Como Sí era la respuesta de interés, los resultados de Minitab fueron los adecuados. Sin embargo, si el orden alfabético de Minitab no da la respuesta de interés, se selecciona cualquier celda de la columna y se usa la secuencia: Editor > Column > Value Order. Minitab le proporcionará la opción de usar un orden especificado por el usuario, pero usted debe poner en segundo lugar de la lista la respuesta de interés en el cuadro define-an-order.

Apéndice 8.2 Estimación por intervalo usando Excel



Lloyd's

A continuación se describe el uso de Excel para calcular intervalos de confianza para la media poblacional y para la proporción poblacional.

Media poblacional: σ conocida

La estimación por intervalo se ilustra empleando el ejemplo de Lloyd´s de la sección 8.1. Se supone que se conoce la desviación estándar poblacional y que es $\sigma = 20$. Las cantidades gastadas por los integrantes de la muestra de tamaño 100 se encuentran en la columna A de la hoja de cálculo de Excel. Para calcular el margen de error para la estimación de la media poblacional se emplean los pasos que se indican a continuación. Se empieza usando la herramienta para estadística descriptiva de Excel, descrita en el capítulo 3.

Paso 1. Seleccionar el menú **Herramientas**
Paso 2. Elegir **Análisis de datos**
Paso 3. Elegir **Estadística descriptiva**
Paso 4. Cuando aparezca el cuadro de diálogo de Estadística descriptiva:
Ingresar A1:A101 en el cuadro **Rango de entrada**
Seleccionar **Agrupado por columnas**
Seleccionar **Rótulos en el primer renglón**
Seleccionar **Rango de salida**
Ingresar C1 en el cuadro **Rango de salida**
Seleccionar **Resumen de estadísticas**
Clic en **OK**

El resumen de estadísticas aparecerá en las columnas C y D. Continúe con el cálculo del margen de error usando la función INTERVALO.CONFIANZA como sigue:

Paso 5. Seleccione la celda C16 e ingrese el título Margen de error
Paso 6. Seleccione la celda D16 e ingrese la fórmula de Excel = INTERVALO.CONFIANZA(0.50,20,100)

Los tres parámetros de esta función son

Alfa = 1 − coeficiente de confianza = 1 − 0.95 = 0.05
La desviación estándar poblacional = 20
El tamaño de la muestra = 100 (*Nota:* Este parámetro aparece como Cuenta en la celda D15.)

La estimación puntual de la media poblacional se encuentra en la celda D3 y el margen de error se encuentra en la celda D16. La estimación puntual de la media poblacional (82) y el margen de error (3.92) permiten calcular fácilmente el intervalo de confianza para la media poblacional.

Media poblacional: σ conocida



NewBalance

La estimación por intervalo se ilustra con los datos de la tabla 8.2 en la que se muestran los saldos en las tarjetas de crédito de 70 hogares. Los datos se encuentran en la columna A de la hoja de cálculo de Excel. Para calcular una estimación puntual y el margen de error de una estimación por intervalo de la media poblacional se siguen los pasos que se indican a continuación. Se emplea la herramienta para estadística descriptiva, vista en el capítulo 3.

Paso 1. Seleccionar el menú **Herramientas**
Paso 2. Elegir **Análisis de datos**
Paso 3. Elegir **Estadística descriptiva**
Paso 4. Cuando aparezca el cuadro de diálogo de Estadística descriptiva:
Ingresar A1:A71 en el cuadro **Rango de entrada**
Seleccionar **Agrupado por columnas**
Seleccionar **Rótulos en el primer renglón**
Seleccionar **Rango de salida**
Ingresar C1 en el cuadro Rango de salida
Seleccionar **Resumen de estadísticas**
Seleccionar **Nivel de confianza para la media**
Ingresar 95 en el cuadro de Nivel de confianza para la media
Clic en **OK**

El resumen estadístico aparecerá en las columnas C y D. La estimación puntual de la media poblacional aparecerá en la celda D3. El margen de error aparecerá como “Nivel de confianza (95.0%)” en la celda D16. La estimación puntual ($9312) y el margen de error ($955) permiten estimar fácilmente el intervalo de confianza para la media poblacional. En la figura 8.10 se muestra el resultado de este procedimiento de Excel.

FIGURA 8.10 ESTIMACIÓN POR INTERVALO DE LA MEDIA POBLACIONAL DE SALDOS EN TARJETAS DE CRÉDITO USANDO EXCEL

	A	B	C	D	E	F
1	**NewBalance**		*NewBalance*			
2	9430					
3	7535		Mean	9312	Estimación puntual	
4	4078		Standard Error	478.9281		
5	5604		Median	9466		
6	5179		Mode	13627		
7	4416		Standard Deviation	4007		
8	10676		Sample Variance	16056048		
9	1627		Kurtosis	−0.296		
10	10112		Skewness	0.18792		
11	6567		Range	18648		
12	13627		Minimum	615		
13	18719		Maximum	19263		
14	14661		Sum	651840		
15	12195		Count	70		
16	10544		Confidence Level(95.0%)	955.4354	Margen de error	
17	13659					
70	9743					
71	10324					
71						

Nota: Los renglones 18 a 69 están ocultos.

Proporción poblacional



Interval p

Esta estimación por intervalo se ilustra usando los datos del estudio de las golfistas presentado en la sección 8.4. Los datos se encuentran en la columna A de la hoja de cálculo de Excel. En los datos recabados, una respuesta Sí implica que la golfista está satisfecha con los horarios de salida disponibles y No, que no es el caso. Excel no proporciona una rutina ya elaborada para la estimación de una proporción poblacional; sin embargo, es fácil elaborar una plantilla para usarla con tal propósito. La plantilla que se muestra en la figura 8.11 proporciona un intervalo de confianza de 95% para la estimación de la proporción de golfistas satisfechas con los horarios de salida disponibles. Observe que en la figura 8.11, en las celdas de la hoja de cálculo que aparece

FIGURA 8.11 PLANTILLA DE EXCEL PARA LA ESTIMACIÓN POR INTERVALO DE UNA PROPORCIÓN POBLACIONAL

	A	B	C	D	E
1	**Response**		**Interval Estimate of a Population Proportion**		
2	Yes				
3	No		**Sample Size**	=COUNTA(A2:A901)	
4	Yes		**Response of Interest**	Yes	
5	Yes		**Count for Response**	=COUNTIF(A2:A901,D4)	
6	No		**Sample Proportion**	=D5/D3	
7	No				
8	No		**Confidence Coefficient**	0.95	
9	Yes		***z* Value**	=NORMSINV(0.5+D8/2)	
10	Yes				
11	Yes		**Standard Error**	=SQRT(D6*(1-D6)/D3)	
12	No		**Margin of Error**	=D9*D11	
13	No				
14	Yes		**Point Estimate**	=D6	
15	No		**Lower Limit**	=D14-D12	
16	No		**Upper Limit**	=D14+D12	
17	Yes				
18	No				
901	Yes				
902					

	A	B	C	D	E	F	G
1	**Response**		**Interval Estimate of a Population Proportion**				
2	Yes						
3	No		**Sample Size**	900			
4	Yes		**Response of Interest**	Yes			
5	Yes		**Count for Response**	396			
6	No		**Sample Proportion**	0.4400			
7	No						
8	No		**Confidence Coefficient**	0.95			
9	Yes		***z* Value**	1.960			
10	Yes						
11	Yes		**Standard Error**	0.0165			
12	No		**Margin of Error**	0.0324			
13	No						
14	Yes		**Point Estimate**	0.4400			
15	No		**Lower Limit**	0.4076			
16	No		**Upper Limit**	0.4724			
17	Yes						
18	No						
901	Yes						
902							

Ingrese la respuesta de interés

Ingrese el coeficiente de confianza

Nota Los renglones 19 a 900 están ocultos.

en segundo plano, se presentan las fórmulas que proporcionan los resultados de la hoja de cálculo que aparece en primer plano. Los pasos para usar la plantilla con este conjunto de datos son los que se dan a continuación.

Paso 1. Ingresar el rango de datos A2:A901 en la fórmula =CONTARA de la celda D3
Paso 2. Ingresar Sí como respuesta de interés en la celda D4
Paso 3. Ingresar el rango de datos A2:A901 en la fórmula =CONTAR.SI de la celda D5
Paso 4. Ingresar 0.95 como coeficiente de confianza en la celda D8.

Esta plantilla proporciona automáticamente el intervalo de confianza en las celdas D15 y D16.

Esta plantilla se usa para calcular un intervalo de confianza para la proporción poblacional en otras aplicaciones. Por ejemplo, para calcular la estimación por intervalo de un nuevo conjunto de datos, se ingresan los nuevos datos muestrales en la columna A de la hoja de cálculo y después se modifican las cuatro celdas que se muestran. Si el nuevo conjunto de datos ya ha sido resumido, no es necesario ingresar los datos muestrales en la hoja de cálculo. En este caso se ingresa el tamaño de la muestra en la celda D3 y la proporción muestral en la celda D6; la plantilla proporcionará el intervalo de confianza para la proporción poblacional. La hoja de cálculo de la figura 8.11 se encuentra en el archivo p del disco compacto que se distribuye con el libro.



CAPÍTULO 9

Prueba de hipótesis

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

JOHN MORRELL & COMPANY*
CINCINNATI, OHIO

John Morrell & Company, que se inició en Inglaterra en 1827, es considerado el fabricante de productos de carne más antiguo de Estados Unidos que ha funcionado con continuidad. Es una subsidiaria, propiedad absoluta aunque independientemente administrada, de Smithfield Foods, Smithfield, Virginia. John Morrell & Company ofrece a los consumidores una amplia línea de productos de carne procesada y carne fresca de puerco de 13 marcas regionales que comprenden John Morrell, E-Z-Cut, Tobin's First Prize, Dinner Bell, Hunter, Kretschmar, Rath, Rodeo, Shenson, Farmers Hickory Brand, Iowa Quality y Peyton's. Cada marca regional disfruta del reconocimiento y la lealtad de sus consumidores.

Las investigaciones de mercado de Morrell proporcionan a los directivos información actualizada acerca de los diversos productos de la empresa y sobre su posición en relación con las otras marcas de productos similares. En un estudio reciente se comparó uno de los productos de Morrell, Beef Pot Roast, con productos similares de dos de los competidores principales. En esta prueba de comparación de los tres productos, se empleó una muestra de consumidores para que indicaran cómo calificaban a los productos en términos de sabor, apariencia, aroma y preferencia.

Una de las cuestiones que se deseaba investigar era si el producto de Morrell era la elección de preferencia de más de 50% de la población de consumidores. Si p representa la proporción poblacional que prefiere el producto de Morrell, la prueba de hipótesis para la cuestión que se investiga es la siguiente:

$$H_0: p \leq 0.50$$
$$H_a: p > 0.50$$

La hipótesis nula H_0 indica que la preferencia por el producto de Morrel es menor o igual que 50%. Si los datos muestrales respaldan el rechazo de H_0 en favor de la hipótesis alternativa H_a, Morrell concluirá que en una comparación de los tres productos, el suyo es preferido por más de 50% de la población de consumidores.

En un estudio independiente se hizo una prueba de degustación empleando una muestra de 224 consumidores de Cincinnati, Milwaukee y Los Ángeles, 150 consumidores eligieron el producto de Morrell como el de su preferencia. A partir del procedimiento estadístico de prueba de hipótesis, se rechazó la hipótesis nula. Mediante el estudio se encontraron evidencias estadísticas que favorecían a H_a y se llegó a la conclusión de que el producto de Morrell es preferido por más de 50% de la población de consumidores.

La estimación puntual de la proporción poblacional es $\bar{p} = 150/224 = 0.67$. De este modo, los datos muestrales sirvieron para hacer publicidad en una revista de alimentos en la cual se mostraba que en una comparación del sabor de los tres productos, el producto de Morrell era "preferido en una relación 2 a 1".

En este capítulo se estudiará cómo formular hipótesis y la forma de elaborar pruebas como la usada por Morrell. Mediante el análisis de datos muestrales se podrá determinar si una hipótesis debe o no rechazarse.



Platillos totalmente listos permiten al consumidor calentarlos y servirlos en la misma charola usada para el microondas. © Cortesía de John Morrel's Convenient Cusine Products.

*Los autores agradecen a Marty Butler vicepresidente de marketing de John Morrell por proporcionarles este artículo para *La estadística en la práctica*.

En los capítulos 7 y 8 se mostró cómo calcular estimaciones puntuales y por intervalo de los parámetros poblacionales. En este capítulo, se continúa con el estudio de la inferencia estadística mostrando la forma de usar la prueba de hipótesis para determinar si una afirmación acerca del valor de un parámetro poblacional debe o no ser rechazada.

Cuando se hace una prueba de hipótesis se empieza por hacer una suposición tentativa acerca del parámetro poblacional. A esta suposición tentativa se le llama **hipótesis nula** y se denota por H_0. Después se define otra hipótesis, llamada **hipótesis alternativa**, que dice lo contrario de lo que establece la hipótesis nula. La hipótesis alternativa se denota H_a.

En el procedimiento de pruebas de hipótesis se usan datos de una muestra para probar dos afirmaciones contrarias indicadas por H_0 y H_a.

En este capítulo se indica el modo de realizar pruebas de hipótesis para medias poblacionales y proporciones poblacionales. Para empezar se facilitan ejemplos que ilustran los métodos para elaborar la hipótesis nula y la hipótesis alternativa.

9.1 Elaboración de las hipótesis nula y alternativa

Para aprender a formular hipótesis correctas se necesita práctica. Al principio hay cierta confusión al elegir H_0 y H_a. Los ejemplos de esta sección muestran varias formas para H_0 y H_a dependiendo de la aplicación.

En algunas aplicaciones no parece obvio cómo formular la hipótesis nula y la hipótesis alternativa. Se debe tener cuidado en estructurar las hipótesis apropiadamente de manera que la conclusión de la prueba de hipótesis proporcione la información que el investigador o la persona encargada de tomar las decisiones desea. Se darán los lineamientos para establecer la hipótesis nula y la hipótesis alternativa en tres tipos de situaciones en las cuales se suele emplear el procedimiento de prueba de hipótesis.

Prueba de una hipótesis de investigación

Considere un determinado modelo de automóvil en el que el rendimiento de la gasolina es 24 millas por galón. Un grupo de investigación elabora un nuevo sistema de inyección de combustible diseñado para dar un mejor rendimiento en millas por galón de gasolina. Para evaluar el nuevo sistema se fabrican varios de éstos, se instalan en los automóviles y se someten a pruebas controladas de manejo. En este caso, el grupo de investigación busca evidencias para concluir que el nuevo sistema *aumenta* la media del rendimiento. La hipótesis de investigación es, entonces, que el nuevo sistema de inyección de combustible proporciona un rendimiento medio mayor a 24 millas por galón de combustible; es decir, $\mu > 24$. Como lineamiento general, una hipótesis de investigación se debe plantear como *hipótesis alternativa*. Por tanto, en este estudio las hipótesis nula y alternativa adecuadas son

$$H_0: \mu \leq 24$$
$$H_a: \mu > 24$$

Se concluye que la hipótesis de investigación es verdadera si los datos muestrales contradicen la hipótesis nula.

Si los resultados obtenidos con la muestra indican que no se puede rechazar H_0, los investigadores no concluirán que el nuevo sistema de inyección de combustible sea mejor. Quizá será necesario continuar investigando y realizar nuevas pruebas. Pero si los resultados muestrales indican que se puede rechazar H_0, los investigadores inferirán que H_a: $\mu > 24$ es verdadera. Esta conclusión proporciona a los investigadores el apoyo estadístico necesario para afirmar que el nuevo sistema aumenta el rendimiento medio en millas por galón. Se considerará la producción del nuevo sistema.

En estudios de investigación como éste, las hipótesis nula y alternativa deben formularse de manera que al rechazar H_0 se apoye la conclusión de la investigación. La hipótesis de la investigación, entonces, debe expresarse como hipótesis alternativa.

Prueba de la validez de una afirmación

Como ilustración de la prueba de la validez de una afirmación, considere una situación en la que un fabricante de refrescos asegura que los envases de dos litros de refresco contienen en promedio, por lo menos, 67.6 onzas de líquido. Se selecciona una muestra de envases de dos litros y se mide su contenido para confirmar lo que asegura el fabricante. En este tipo de situaciones de prueba de hipótesis, se suele suponer que el dicho del fabricante es verdad a menos que las evidencias muestrales indiquen lo contrario. Si se sigue este método en el ejemplo de los refrescos, las hipótesis nula y alternativa se establecen como sigue.

$$H_0: \mu \geq 67.6$$
$$H_a: \mu < 67.6$$

Si los resultados muestrales indican que no se puede rechazar H_0, entonces no se cuestiona lo que asegura el fabricante. Pero si los resultados muestrales indican que se puede rechazar H_0, lo que se inferirá es que H_a: $\mu < 67.6$ es verdad. Si tal es la conclusión, las evidencias estadísticas indican que el dicho del fabricante no es correcto y que los envases de refrescos contienen en promedio menos de las 67.6 onzas que se asegura contienen. Se considerará realizar las acciones correspondientes en contra del fabricante.

A lo que asegura un fabricante se le suele dar el beneficio de la duda y se establece como hipótesis nula. Si se rechaza la hipótesis nula se concluye que su dicho es falso.

En toda situación en la que se desee probar la validez de una afirmación, la hipótesis nula se suele basar en la suposición de que la afirmación sea verdadera. Entonces, la hipótesis alternativa se formula de manera que rechazar H_0 proporcione la evidencia estadística de que la suposición establecida es incorrecta. Siempre que se rechace H_0 deberán considerarse las medidas necesarias para corregir la afirmación.

Prueba en situaciones de toma de decisión

Cuando se prueba una hipótesis de investigación o la validez de una afirmación, se toman medidas si se rechaza H_0; sin embargo, en algunas situaciones se toman tanto si no se puede rechazar H_0 como si se puede rechazar H_0. En general, este tipo de situaciones se presentan cuando la persona que debe tomar una decisión tiene que elegir entre dos líneas de acción, una relacionada con la hipótesis nula y otra con la hipótesis alternativa. Por ejemplo, con base en una muestra de las piezas de un pedido recibido, el inspector de control de calidad tiene que decidir si acepta el pedido o si lo regresa al proveedor debido a que no satisface las especificaciones. Suponga que una especificación para unas piezas determinadas sea que su longitud media deba ser de dos pulgadas. Si la longitud media es menor o mayor a dos pulgadas, las piezas ocasionarán problemas de calidad en la operación de ensamblado. En este caso, las hipótesis nula y alternativa se formulan como sigue.

Este tipo de prueba de hipótesis se emplea en el procedimiento de control de calidad conocido como muestreo de aceptación de lotes.

$$H_0: \mu = 2$$
$$H_a: \mu \neq 2$$

Si los resultados muestrales indican que no se puede rechazar H_0, el inspector de control de calidad no tendrá razón para dudar que el pedido satisfaga las especificaciones y aceptará el pedido. Pero si los resultados muestrales indican que H_0 se debe rechazar, se concluirá que las piezas no satisfacen las especificaciones. En este caso, el inspector de control de calidad tendrá evidencias suficientes para regresar el pedido al proveedor. Así, se ve que en este tipo de situaciones, se toman medidas en ambos casos, cuando H_0 no se puede rechazar y cuando H_0 se puede rechazar.

Resumen de las formas para las hipótesis nula y alternativa

Las pruebas de hipótesis de este capítulo se refieren a dos parámetros poblacionales: la media poblacional y la proporción poblacional. A partir de la situación, las pruebas de hipótesis para un parámetro poblacional asumen una de estas tres formas: en dos se emplean desigualdades en la hipótesis nula y en la tercera se aplica una igualdad en la hipótesis nula. En las pruebas de hipótesis para la media poblacional, μ_0 denota el valor hipotético y para la prueba de hipótesis hay que escoger una de las formas siguientes.

Aquí se muestran las tres formas que pueden tener H_0 y H_a. Observe que en la hipótesis nula H_0 siempre aparece la igualdad.

$$H_0: \mu \geq \mu_0 \qquad H_0: \mu \leq \mu_0 \qquad H_0: \mu = \mu_0$$
$$H_a: \mu < \mu_0 \qquad H_a: \mu > \mu_0 \qquad H_a: \mu \neq \mu_0$$

Por razones que serán claras más tarde, a las dos primeras formas se les llama pruebas de una cola. A la tercera se le llama prueba de dos colas.

Con frecuencia se tienen situaciones en las que no es obvio cómo elegir H_0 y H_a y se debe tener cuidado para elegirlas en forma adecuada. Sin embargo, como se observa en las formas anteriores, la igualdad (ya sea $\geq$, $\leq$, o $=$) debe aparecer *siempre* en la hipótesis nula. Al elegir la

forma adecuada para H_0 y H_a hay que tener en mente que la hipótesis alternativa a menudo es lo que la prueba está tratando de demostrar. Por tanto, preguntarse si el usuario está buscando evidencias en apoyo de $\mu < \mu_0$, $\mu > \mu_0$ o $\mu \neq \mu_0$ ayuda a determinar H_a. Los ejercicios siguientes tienen por objeto aportar práctica en la elección de la forma adecuada de una prueba de hipótesis para la media poblacional

Ejercicios

1. El gerente de Danvers-Hilton Resort afirma que la cantidad media que gastan los huéspedes en un fin de semana es de $600 o menos. Un miembro del equipo de contadores observó que en los últimos meses habían aumentado tales cantidades. El contador emplea una muestra de cuentas de fin de semana para probar la afirmación del gerente.
   a. ¿Qué forma de hipótesis deberá usar para probar la afirmación del gerente? Explique.

$$H_0\colon \mu \geq 600 \qquad H_0\colon \mu \leq 600 \qquad H_0\colon \mu = 600$$
$$H_a\colon \mu < 600 \qquad H_a\colon \mu > 600 \qquad H_a\colon \mu \neq 600$$

   b. ¿Cuál es la conclusión apropiada cuando no se puede rechazar la hipótesis nula H_0?
   c. ¿Cuál es la conclusión apropiada cuando se puede rechazar la hipótesis nula H_0?



2. El gerente de un negocio de venta de automóviles está pensando en un nuevo plan de bonificaciones, con objeto de incrementar el volumen de ventas. Al presente, el volumen medio de ventas es 14 automóviles por mes. El gerente desea realizar un estudio para ver si el plan de bonificaciones incrementa el volumen de ventas. Para recolectar los datos una muestra de vendedores venderá durante un mes bajo el nuevo plan de bonificaciones.
   a. Dé las hipótesis nula y alternativa más adecuadas para este estudio.
   b. Comente la conclusión resultante en el caso en que H_0 no pueda rechazarse.
   c. Comente la conclusión que se obtendrá si H_0 puede rechazarse.
3. Una línea de operación está diseñada para llenar empaques de 32 onzas de detergente para lavar. Con periodicidad se selecciona una muestra de los empaques y se pesan para determinar si no se están llenando con un peso mayor o menor del indicado. Si los datos muestrales llevan a la conclusión de que hay exceso o falta de llenado, se suspende la producción y se ajusta al llenado correcto.
   a. Formule las hipótesis nula y alternativa que ayudarán a determinar si se debe detener la producción y ajustar el peso.
   b. Comente sobre la conclusión y la decisión en caso en que H_0 no se pueda rechazar.
   c. Comente acerca de la conclusión y la decisión en caso en que H_0 se pueda rechazar.
4. Debido a los costos y al tiempo de adaptación de la producción, un director de fabricación antes de implantar un nuevo método de fabricación, debe convencer al gerente de que ese nuevo método de fabricación reducirá los costos. El costo medio del actual método de producción es $220 por hora. En un estudio se medirá el costo del nuevo método durante un periodo muestral de producción,
   a. Dé las hipótesis nula y alternativa más adecuadas para este estudio.
   b. Haga un comentario sobre la conclusión cuando H_0 no pueda rechazarse.
   c. Dé un comentario sobre la conclusión cuando H_0 pueda rechazarse.



Errores tipo I y II

Las hipótesis nula y alternativa son afirmaciones opuestas acerca de la población. Una de las dos, ya sea la hipótesis nula o la alternativa es verdadera, pero no ambas. Lo ideal es que la prueba de hipótesis lleve a la aceptación de H_0 cuando H_0 sea verdadera y al rechazo de H_0 cuando H_a

TABLA 9.1 ERRORES Y CONCLUSIONES CORRECTAS EN LAS PRUEBAS DE HIPÓTESIS

		Situación en la población	
		H_0 verdadera	H_a verdadera
Conclusión	Se acepta H_0	Conclusión correcta	Error tipo II
	Se rechaza H_0	Error tipo I	Conclusión correcta

sea verdadera. Por desgracia, las conclusiones correctas no siempre son posibles. Como la prueba de hipótesis se basa en una información muestral debe tenerse en cuenta que existe la posibilidad de error. La tabla 9.1 ilustra las dos clases de errores comunes en una prueba de hipótesis.

En el primer renglón de la tabla 9.1 se muestra lo que sucede cuando se acepta H_0. Si H_0 es verdadera la conclusión es correcta. Pero, si H_a es verdadera se comete un **error tipo II**; es decir, se acepta H_0 cuando es falsa. En el segundo renglón de la tabla 9.1 se muestra lo que sucede si la conclusión es rechazar H_0. Si H_0 es verdadera se comete un **error tipo I**; es decir, se rechaza H_0 cuando es verdadera. Pero si H_a es verdadera, es correcto rechazar H_0.

Recuerde la ilustración de la prueba de hipótesis vista en la sección 9.1 en la cual un grupo de investigación elaboraba un nuevo sistema de inyección de combustible con objeto de aumentar el rendimiento combustible en un determinado modelo de automóvil. Como con el sistema actual el rendimiento promedio es 24 millas por galón, la prueba de hipótesis se formuló como sigue.

$$H_0: \mu \leq 24$$
$$H_a: \mu > 24$$

La hipótesis alternativa, H_a: $\mu > 24$, indica que los investigadores buscan evidencias muestrales que apoyen la conclusión de que con el nuevo sistema de inyección de combustible la media poblacional del rendimiento es mayor que 24 millas por galón.

En esta aplicación, el error tipo I de rechazar H_0 cuando es verdadera corresponde a la afirmación de los investigadores de que el nuevo sistema mejora el rendimiento ($\mu > 24$) cuando en realidad el nuevo sistema no es nada mejor que el actual. En cambio, el error tipo II de aceptar H_0 cuando es falsa corresponde a la conclusión de los investigadores de que el nuevo sistema no es mejor que el actual ($\mu \leq 24$) cuando en realidad el nuevo sistema sí mejora el rendimiento.

En esta prueba de hipótesis del rendimiento, la hipótesis nula es H_0: $\mu \leq 24$. Admita que la hipótesis nula es verdadera como una igualdad; es decir $\mu = 24$. A la probabilidad de cometer un error tipo I cuando la hipótesis nula es verdadera como igualdad se le conoce como **nivel de significancia**. Por tanto en la prueba de hipótesis del rendimiento de combustible, el nivel de significancia es la probabilidad de rechazar H_0: $\mu \leq 24$ cuando $\mu = 24$. Dada la importancia de este concepto se redacta otra vez la definición de nivel de significancia.

NIVEL DE SIGNIFICANCIA

El nivel de significancia es la probabilidad de cometer un error tipo I cuado la hipótesis nula es verdadera como igualdad.

Para denotar el nivel de significancia se usa la letra griega α (alfa), y los valores que se suelen usar para α son 0.05 y 0.01.

En la práctica la persona responsable de la prueba de hipótesis especifica el nivel de significancia. Al elegir α se controla la probabilidad de cometer un error tipo I. Si el costo de cometer un error tipo I es elevado, los valores pequeños de α son preferibles. Si el costo de cometer un error tipo I no es demasiado elevado, entonces se usan valores mayores para α. A las aplicaciones de la prueba de hipótesis en que sólo se controla el error tipo I se les llama *pruebas de significancia.* Muchas aplicaciones de las pruebas de hipótesis son de este tipo.

Si los datos muestrales son consistentes con la hipótesis nula H_0, se dirá que "no se rechaza H_0". Se prefiere esta conclusión a la conclusión "se acepta H_0" porque con la conclusión de aceptar H_0 se corre el riesgo de cometer un error tipo II.

Aunque en la mayor parte de las aplicaciones de las pruebas de hipótesis se controla la probabilidad de cometer un error tipo I, no siempre sucede lo mismo con un error tipo II. Por tanto, si se decide aceptar H_0 no es posible establecer la confianza en esa decisión. Debido a la incertidumbre de cometer un error tipo II al realizar una prueba de significancia los dedicados a la estadística suelen recomendar que se diga "no se rechaza H_0" en lugar de "se acepta H_0". Decir "no se rechaza H_0" implica la recomendación de reservarse tanto el juicio como la acción. En efecto al no aceptar directamente H_0, se evita el riesgo de cometer un error tipo II. Siempre que no se determine y controle la probabilidad de cometer un error tipo II, no se dirá "se acepta H_0". En esos casos sólo son posibles dos conclusiones: *no se rechaza* H_0 o *se rechaza* H_0.

Aunque controlar el error tipo II en una prueba de hipótesis es poco común, es posible. En las secciones 9.7 y 9.8 se ilustra el procedimiento para controlar y determinar la probabilidad de cometer un error tipo II. Si se ha establecido un control adecuado de este error, las medidas basadas en la conclusión "se acepta H_0" son adecuadas.

NOTAS Y COMENTARIOS

Walter Williams, columnista y profesor de economía en la universidad George Mason indica que siempre existe la posibilidad de cometer un error tipo I o un error tipo II al tomar una decisión (*The Cincinnati Enquirer*, 14 de agosto de 2005). Hace notar que la Food and Drug Administration corre el riesgo de cometer estos errores en sus procedimientos para la aprobación de medicamentos. Cuando comete un error tipo I, la FDA no aprueba un medicamento que es seguro y efectivo. Al cometer un error tipo II, la FDA aprueba un medicamento que presenta efectos secundarios imprevistos. Sin importar la decisión que se tome, la probabilidad de cometer un error costoso no se puede eliminar.

Ejercicios



5. Nielsen informó que los hombres jóvenes estadounidenses ven diariamente 56.2 minutos de televisión en las horas de mayor audiencia (*The Wall Street Journal Europe*, 18 de noviembre de 2003). Un investigador cree que en Alemania, los hombres jóvenes ven más tiempo la televisión en las horas de mayor audiencia. Este investigador toma una muestra de hombres jóvenes alemanes y registra el tiempo que ven televisión en un día. Los resultados muestrales se usan para probar las siguientes hipótesis nula y alternativa.

$$H_0\colon \mu \leq 56.2$$
$$H_a\colon \mu > 56.2$$

   a. En esta situación, ¿cuál es el error tipo I? ¿Qué consecuencia tiene cometer este error?
   b. En esta situación, ¿cuál es el error tipo II? ¿Qué consecuencia tiene cometer este error?

6. En la etiqueta de una botella de jugo de naranja de 3 cuartos (de galón) dice que el jugo de naranja contiene en promedio 1 gramo o menos de grasa. Responda a las preguntas siguientes relacionadas con una prueba de hipótesis para probar lo que dice en la etiqueta.
   a. Dé las hipótesis nula y alternativa adecuadas.

b. En esta situación, ¿cuál es el error tipo I? ¿Qué consecuencia tiene cometer este error?
c. En esta situación, ¿cuál es el error tipo II? ¿Qué consecuencia tiene cometer este error?

7. El personal de ventas de Carpetland vende, en promedio, $8000 semanales. Steve Contois, vicepresidente de la empresa, propone un plan de compensaciones con nuevos incentivos de venta. Steve espera que los resultados de un periodo de prueba le permitirán concluir que el plan de compensaciones aumenta el promedio de ventas de los vendedores.
   a. Dé las hipótesis nula y alternativa adecuadas.
   b. En esta situación, ¿cuál es el error tipo I? ¿Qué consecuencia tiene cometer este error?
   c. En esta situación, ¿cuál es el error tipo II? ¿Qué consecuencia tiene cometer este error?

8. Suponga que se va a implantar un nuevo método de producción si mediante una prueba de hipótesis se confirma la conclusión de que el nuevo método de producción reduce el costo medio de operación por hora.
   a. Dé las hipótesis nula y alternativa adecuadas si el costo medio de producción actual por hora es $220.
   b. En esta situación, ¿cuál es el error tipo I? ¿Qué consecuencia tiene cometer este error?
   c. En esta situación, ¿cuál es el error tipo II? ¿Qué consecuencia tiene cometer este error?



Media poblacional: σ conocida

En el capítulo 8 se dijo que el caso σ conocida se refiere a aplicaciones en las que se cuenta con datos históricos o con alguna información que permita obtener buenas estimaciones de la desviación estándar poblacional antes de tomar la muestra. En tales casos, para propósitos prácticos, se considera que se conoce la desviación estándar poblacional. En esta sección se muestra cómo realizar una prueba de hipótesis para la media poblacional en el caso en que σ es conocida.

Los métodos que se presentan en esta sección dan resultados exactos si la población de la que se selecciona la muestra tiene distribución normal. En los casos en los que no sea razonable suponer que la población tiene una distribución normal, se pueden aplicar estos métodos siempre y cuando el tamaño de la muestra sea suficientemente grande. Al final de esta sección se proporcionan algunos consejos prácticos en relación con la distribución de la población.

Prueba de una cola

Una **prueba de una cola** para la media poblacional tiene una de las dos formas siguientes.

Prueba de la cola inferior (o izquierda)	**Prueba de la cola superior (o derecha)**
$H_0\text{: } \mu \geq \mu_0$	$H_0\text{: } \mu \leq \mu_0$
$H_a\text{: } \mu < \mu_0$	$H_a\text{: } \mu > \mu_0$

A continuación se presenta un ejemplo de una prueba para la cola inferior.

La Federal Trade Commission, FTC, realiza periódicamente estudios estadísticos con objeto de comprobar las afirmaciones de los fabricantes acerca de sus productos. Por ejemplo, en la etiqueta de una lata grande de Hilltop Coffee dice que la lata contiene 3 libras de café. La FTC sabe que el proceso de producción de Hilltop no permite llenar las latas con 3 libras exactas de café por lata, incluso si la media poblacional del peso de llenado de todas las latas es de 3 libras por lata. Sin embargo, mientras la media poblacional del peso de llenado sea por lo menos 3 libras por lata, los derechos del consumidor estarán protegidos. Por tanto, la FTC interpreta que la información de la etiqueta de una lata grande de café Hilltop tiene una media poblacional del peso de llenado de por lo menos 3 libras por lata. Se mostrará cómo verificar esto realizando una prueba de hipótesis de la cola inferior.

El primer paso es dar las hipótesis nula y alternativa para la prueba. Si la media poblacional del peso de llenado es por lo menos 3 libras por lata, lo que afirma Hilltop correcto. Esto establece la hipótesis nula de la prueba. No obstante, si la media poblacional del peso de llenado es

menor que 3 libras por lata, la afirmación de Hilltop es incorrecta. Así, se establece la hipótesis alternativa. Si μ denota la media poblacional del peso de llenado, las hipótesis nula y alternativa son las siguientes:

$$H_0: \mu \geq 3$$
$$H_a: \mu < 3$$

Observe que el valor hipotético de la media poblacional es $\mu_0 = 3$.

Si los datos muestrales indican que H_0 no se puede rechazar, las evidencias estadísticas no conducirán a concluir que ha habido una violación en lo que se afirma en la etiqueta. Luego, no habrá ninguna acción en contra de Hilltop. Pero, si los datos muestrales indican que se puede rechazar H_0 se concluirá que la hipótesis alternativa H_a: $\mu < 3$ es verdadera. En este caso la conclusión de que hay falta de peso y un cargo por violación a lo que se establece en la etiqueta estará justificada.

Suponga que se selecciona una muestra de 36 latas de café y se calcula la media muestral $\bar{x}$ como una estimación de la media poblacional μ. Si el valor de la media muestral $\bar{x}$ es menor que 3 libras, los resultados muestrales despertarán dudas sobre lo que establece la hipótesis nula. Lo que se busca saber es cuánto menor que 3 libras tiene que ser $\bar{x}$ para declarar que la diferencia es significativa y se esté dispuesto a correr el riesgo de cometer un error tipo I al acusar indebidamente a Hilltop de una violación de lo que establece en la etiqueta. Aquí el factor clave es el valor elegido como nivel de significancia por quien tomará la decisión.

Como ya se dijo, el nivel de significancia, que se denota α, es la probabilidad de cometer un error tipo I al rechazar la hipótesis nula cuando ésta, considerada en forma de una igualdad, es verdadera. La persona que tomará la decisión debe especificar el nivel de significancia. Si el costo de cometer un error tipo I es alto, se deberá elegir un valor pequeño para el nivel de significancia. Si el costo no es alto, se deberá elegir un valor grande, es lo más apropiado. En el caso del café Hilltop, el director del programa de pruebas de la FTC afirma: "Si la empresa satisface sus especificaciones de peso en $\mu = 3$, no tomaré ninguna medida en su contra. Pero, estoy dispuesto a asumir un riesgo de 1% de cometer tal error". De acuerdo con lo dicho por el director, el nivel de significancia en esta prueba de hipótesis se establece en $\alpha = 0.01$. Así, la prueba de hipótesis deberá diseñarse de manera que la probabilidad de cometer un error tipo I cuando $\mu = 3$ sea 0.01.

En este estudio sobre Hilltop Coffee, al dar las hipótesis nula y alternativa y especificar el nivel de significancia para la prueba, se han dado los dos primeros pasos requeridos en cualquier prueba de hipótesis. Con esto se está listo para el tercer paso en una prueba de hipótesis: recoger los datos muestrales y calcular el valor de lo que se conoce como el estadístico de prueba.

Estadístico de prueba. En el estudio de Hilltop Coffee, las pruebas realizadas con anterioridad por la FTC muestran que la desviación estándar puede considerarse conocida, siendo su valor $\sigma = 0.18$. Estas pruebas muestran, también, que puede considerarse que la población de los pesos de llenado tiene una distribución normal. Según lo visto en el capítulo 7 sobre distribuciones muestrales, se sabe que si la población de la que se toma la muestra tiene una distribución normal, la distribución muestral de $\bar{x}$ también es normal. En consecuencia, en el estudio de Hilltop Coffee, la distribución muestral de $\bar{x}$ será una distribución normal. Con un valor conocido de $\sigma = 0.18$ y un tamaño de muestra de $n = 36$, en la figura 9.1 se muestra la distribución muestral de $\bar{x}$ si la hipótesis nula, considerada como igualdad, es verdadera, es decir, cuando $\mu = \mu_0 = 3$.* Observe que el error estándar de $\bar{x}$ está dado por $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.18/\sqrt{36} = 0.03$.

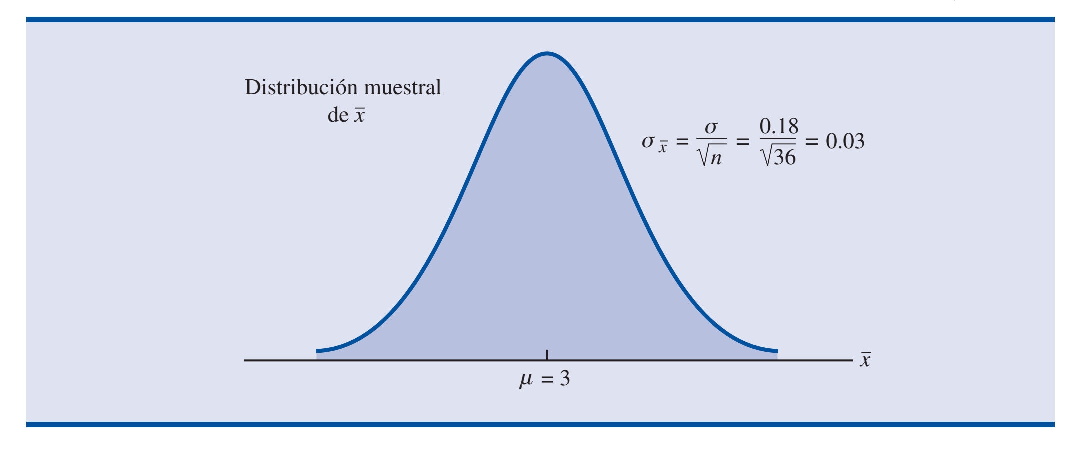
El error estándar de $\bar{x}$ es la desviación estándar de la distribución muestral de $\bar{x}$.

Como la distribución muestral de $\bar{x}$ está distribuida normalmente, la distribución muestral de

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - 3}{0.03}$$

*Cuando se elaboran distribuciones muestrales para una prueba de hipótesis, se supone que H_0 se satisface como igualdad.

FIGURA 9.1 DISTRIBUCIÓN MUESTRAL DE $\bar{x}$ EN EL ESTUDIO DE HILLTOP COFFEE SIENDO LA HIPÓTESIS NULA VERDADERA COMO IGUALDAD ($\mu = 3$)



es una distribución normal estándar. Si el valor de $z = -1$, esto significa que el valor de $\bar{x}$ es una desviación estándar menor que el valor hipotético de la media; si el valor de $z = -2$, esto significa que el valor de $\bar{x}$ es dos errores estándar menor que el valor hipotético de la media y así sucesivamente. Para hallar la probabilidad que corresponde a cualquier valor de z en la cola inferior se usa la tabla de probabilidad normal estándar. Por ejemplo, el área en la cola inferior para $z = -3.00$ es 0.0013. Así, la probabilidad de obtener un valor de z que sea tres o más errores estándar menor que la media es 0.0013. Como resultado, la probabilidad de obtener un valor de $\bar{x}$ que sea 3 o más errores estándar menor que la media poblacional hipotética $\mu_0 = 3$ también es 0.0013. Si la hipótesis nula es verdadera, un resultado así es poco probable.

En una prueba de hipótesis para la media poblacional en el caso σ conocida, se emplea la variable aleatoria normal estándar z como **estadístico de prueba** para determinar si $\bar{x}$ se desvía lo suficiente del valor hipotético de μ como para justificar el rechazo de la hipótesis nula. Como $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, el estadístico de prueba es el siguiente.

ESTADÍSTICO DE PRUEBA EN UNA PRUEBA DE HIPÓTESIS PARA LA MEDIA POBLACIONAL σ CONOCIDA

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{9.1}$$

La cuestión clave en una prueba de la cola inferior es: ¿Qué tan pequeño debe ser el estadístico de prueba z para que se decida rechazar la hipótesis nula? Para responder esta pregunta se usan dos métodos.

Método del valor *p*. En el método del valor-p se usa el valor del estadístico de prueba z para calcular una probabilidad llamada **valor-*p***.

Si el valor-p es pequeño, esto indica que el valor del estadístico de prueba es inusual bajo la suposición de que H_0 es verdadera.

VALOR-p

Un valor-p es una probabilidad que aporta una medida de una evidencia suministrada por la muestra contra la hipótesis nula. Valores-p pequeños indican una evidencia mayor contra la hipótesis nula.

El valor-p se usa para determinar si la hipótesis nula debe ser rechazada.

Ahora se verá cómo se calcula y usa el valor-*p*. Para calcular el valor-*p* se usa el valor del estadístico de prueba. El método a seguir depende de si se trata de una prueba de la cola inferior, de la cola superior o de dos colas. En una prueba de la cola inferior, el valor-*p* es la probabilidad de conseguir un valor del estadístico de prueba tan pequeño o menor que el obtenido con la muestra. Por ende, para calcular el valor-*p* en una prueba de la cola inferior, en el caso σ conocida, se halla el área bajo la curva normal estándar a la izquierda del estadístico de prueba. Una vez calculado el valor-*p* se decide si es lo suficientemente pequeño para rechazar la hipótesis nula; como se verá más adelante, para esta decisión hay que comparar el valor-*p* con el nivel de significancia.



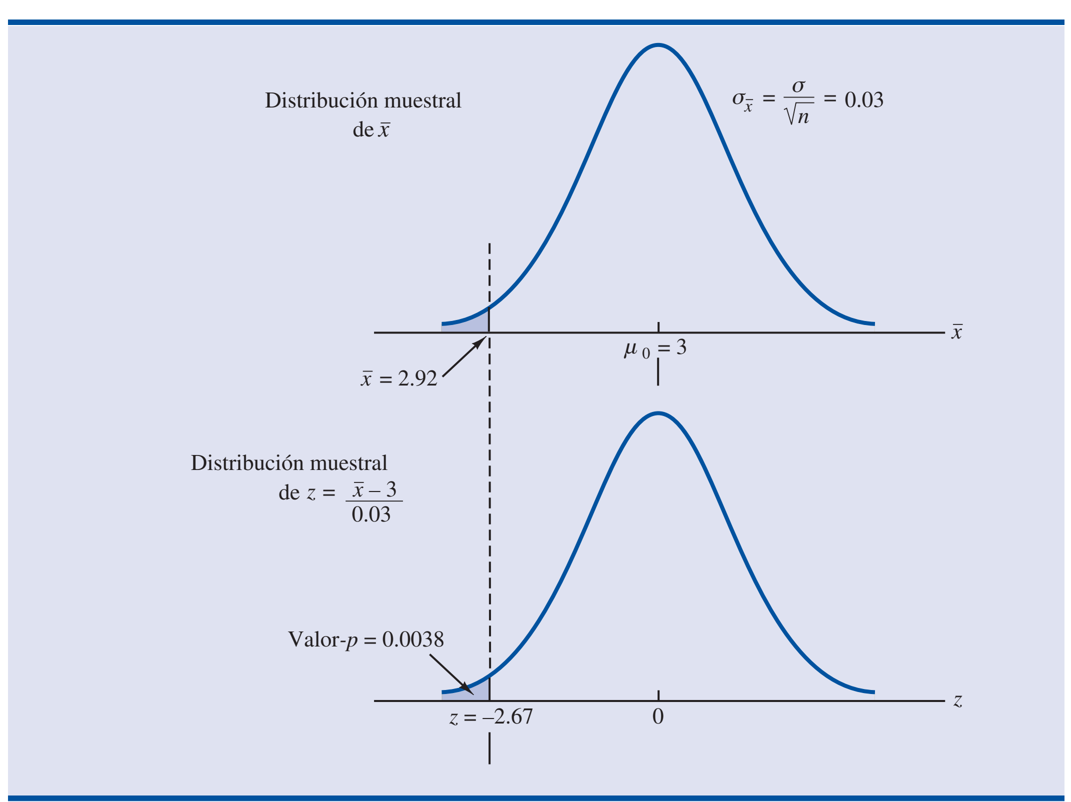
Ahora se calcula el valor-*p* para la prueba de la cola inferior del estudio de Hilltop Coffee. Suponga que en la muestra de las 36 latas de café, la media obtenida es $\bar{x} = 2.92$ libras. ¿Es $\bar{x} = 2.92$ lo suficientemente pequeña para que se rechace la hipótesis nula? Como es una prueba de la cola inferior, el valor-*p* es el área bajo la curva normal estándar a la izquierda del estadístico de prueba. Usando $\bar{x} = 2.92$, $\sigma = 0.18$ y $n = 36$, se calcula el valor del estadístico de prueba z.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{2.92 - 3}{0.18/\sqrt{36}} = -2.67$$

Por consiguiente, el valor-*p* es la probabilidad de que el estadístico de prueba z sea menor o igual que -2.76 (el área bajo la curva normal estándar a la izquierda del estadístico de prueba).

En la tabla de probabilidad normal estándar aparece que el área en la cola inferior para $z = -2.67$ es 0.0038. En la figura 9.2 se muestra que a $\bar{x} = 2.92$ le corresponde $z = -2.67$ y el valor-$p = 0.0038$. Este valor-*p* indica que si la muestra se ha tomado de una población con $\mu = 3$, la probabilidad de obtener una media muestral $\bar{x} = 2.92$ o menor (y un estadístico de prueba de -2.67) es pequeña. Este valor-*p* no favorece mucho a la hipótesis nula, pero, ¿es lo suficiente-

FIGURA 9.2 VALOR-*p* EN EL ESTUDIO DE HILLTOP COFFEE, EN EL QUE $\bar{x} = 2.92$ Y $z = -2.67$



mente pequeño como para que se rechace H_0? La respuesta depende del nivel de significancia de la prueba.

Como se indicó antes, el director del programa de pruebas de la FTC eligió como nivel de significancia 0.01. Elegir $\alpha = 0.01$ significa que él está dispuesto a tolerar una probabilidad de 0.01 de rechazar la hipótesis nula cuando sea verdadera como igualdad ($\mu_0 = 3$). La muestra de 36 latas de Hilltop Coffee dio como resultado un valor-$p = 0.0038$, lo que significa que la probabilidad de obtener $\bar{x} = 2.92$ o menor, si la hipótesis nula es verdadera considerada como igualdad, es 0.0038. Como 0.0038 es menor que $\alpha = 0.01$, se rechaza H_0. De manera que para el nivel de significancia 0.01 se encontraron evidencias estadísticas suficientes para rechazar la hipótesis nula.

Ahora ya se puede establecer la regla general para determinar cuándo rechazar la hipótesis nula al usar el método del valor-p. Dado un nivel de significancia α, la regla para el rechazo usando el método del valor-p es la siguiente:

REGLA PARA EL RECHAZO USANDO EL VALOR-p

$$\text{Rechazar } H_0 \text{ si el valor-}p \leq \alpha$$

En la prueba para Hilltop Coffee, el valor-$p = 0.0038$ hizo que se rechazara la hipótesis nula. Aunque la base para tomar la decisión de rechazar la hipótesis nula fue comparar el valor-p con el nivel de significancia especificado por el director de la FTC, el valor-p encontrado indica que H_0 se hubiera rechazado para cualquier valor $\alpha \geq .0038$. Debido a esto, el valor-p se conoce también como *nivel de significancia observado.*

Diferentes personas tienen opiniones distintas respecto del costo de cometer un error tipo I y pueden elegir niveles de significancia distintos. Al proporcionar el valor-p como parte de los resultados de la prueba de hipótesis, quizás otra persona compare el valor-p con su propio nivel de significancia y tome otra decisión respecto de rechazar o no H_0.

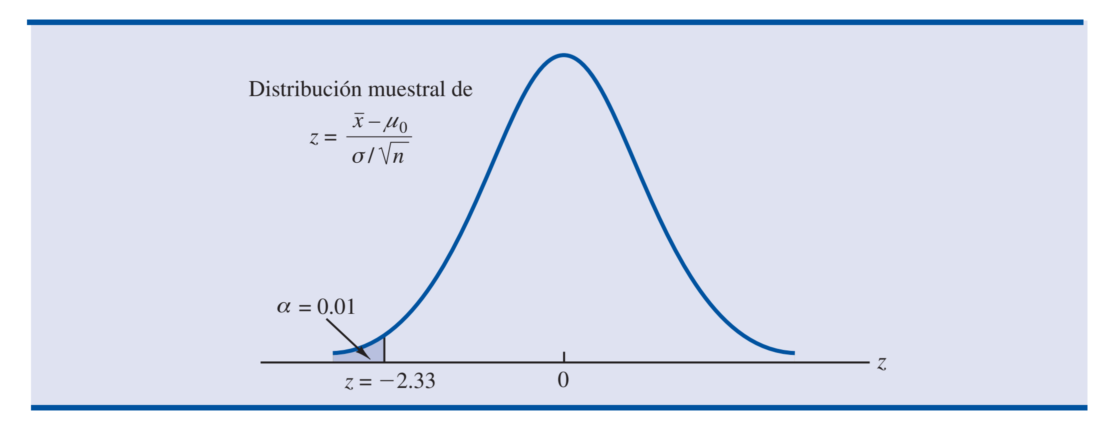
Método del valor crítico. En el método del valor crítico primero se determina un valor para el estadístico de prueba llamado **valor crítico**. En una prueba de la cola inferior, el valor crítico sirve como punto de referencia para determinar si el valor del estadístico de prueba es lo suficientemente pequeño para rechazar la hipótesis nula. El valor crítico es el valor del estadístico de prueba que corresponde a un área α (nivel de significancia) en la cola inferior de la distribución muestral del estadístico de prueba. En otras palabras, el valor crítico es el mayor valor del estadístico de prueba que hará que se rechace la hipótesis nula. A continuación, de nuevo con el ejemplo de Hilltop Coffee, se verá cómo funciona este método.

En el caso σ conocida, la distribución muestral del estadístico de prueba z es la distribución normal estándar. Por tanto, el valor crítico es el valor del estadístico de prueba que corresponde a un área $\alpha = 0.01$ en la cola inferior de la distribución normal estándar. En la tabla de probabilidad normal estándar aparece que $z = -2.33$ proporciona un área de 0.01 en la cola inferior (véase figura 9.3). De manera que si con la muestra se obtiene un valor del estadístico de prueba menor o igual a -2.33, el valor-p correspondiente será menor o igual a 0.01; en este caso se deberá rechazar la hipótesis nula. Entonces, en el estudio de Hilltop Coffee, la regla para el rechazo usando el valor crítico para un nivel de significancia de 0.01 es

$$\text{Rechazar } H_0 \text{ si } z \leq -2.33$$

En el ejemplo del Hilltop Coffee, $\bar{x} = 2.92$ y el estadístico de prueba es $z = -2.67$. Como $z = -2.67 < -2.33$, se puede rechazar H_0 y concluir que Hilltop Coffee está llenando las latas con menos peso del debido.

FIGURA 9.3 VALOR CRÍTICO = −2.33 EN LA PRUEBA DE HIPÓTESIS DE HILLTOP COFFEE



La regla de rechazo se puede generalizar empleando el método del valor crítico para cualquier nivel de significancia. La regla de rechazo en una prueba de la cola inferior es:

REGLA PARA EL RECHAZO EN UNA PRUEBA DE LA COLA INFERIOR: MÉTODO DEL VALOR CRÍTICO

$$\text{Rechazar } H_0 \text{ si } z \leq -z_\alpha$$

donde $-z_\alpha$ es el valor crítico; es decir, el valor z que proporciona un área α en la cola inferior de la distribución normal estándar.

En las pruebas de hipótesis, el método del valor-p y el método del valor crítico llevarán siempre a la misma decisión; siempre que el valor-p sea menor o igual que α, el valor del estadístico de prueba será menor o igual al valor crítico. La ventaja del método del valor-p es que dice *cuán* significativos son los resultados (el nivel de significancia observado). Si se usa el método del valor crítico sólo se sabe que los resultados son significativos al nivel de significancia establecido.

Al principio de esta sección se dijo que las pruebas de una cola, para la media poblacional, toman una de las dos formas siguientes:

Prueba de la cola inferior	**Prueba de la cola superior**
$H_0\colon\ \mu \geq \mu_0$	$H_0\colon\ \mu \leq \mu_0$
$H_a\colon\ \mu < \mu_0$	$H_a\colon\ \mu > \mu_0$

El estudio de Hilltop Coffee se usó para ilustrar cómo realizar una prueba de la cola inferior. El mismo método general se usa para realizar una prueba de la cola superior. Para esta prueba también se calcula el estadístico de prueba z usando la ecuación (9.1). Pero, en una prueba de la cola superior, el valor-p es la probabilidad de obtener un valor para el estadístico de prueba tan grande o mayor que el obtenido con la muestra. Por tanto, para calcular el valor-p para una prueba de la cola superior en el caso de σ conocida, es necesario hallar el área bajo la curva normal estándar a la derecha del estadístico de prueba. Usando el método del valor crítico la hipótesis nula se rechaza si el valor del estadístico de prueba es mayor o igual al valor crítico z_α; en otras palabras, se rechaza H_0 si $z \geq z_\alpha$.

Prueba de dos colas

En las pruebas de hipótesis la forma general de una **prueba de dos colas** es la siguiente:

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

En esta subsección se muestra cómo realizar una prueba de dos colas para la media poblacional en el caso en que se conoce σ. Como ilustración se considera el caso de una prueba de hipótesis en la empresa MaxFlight

La U.S. Golf Association, USGA, establece reglas que deben satisfacer los fabricantes de equipos de golf si quieren que sus productos se acepten en los eventos de USGA. MaxFlight emplea procesos de fabricación de alta tecnología para producir pelotas de golf que tienen una distancia media de recorrido de 295 yardas. Sin embargo, algunas veces el proceso se desajusta y se producen pelotas de golf que tienen una distancia media de recorrido diferente a 295 yardas. Cuando la distancia media es menor que 295 yardas, a la empresa le preocupa perder clientes porque las pelotas de golf no proporcionen la distancia anunciada. Cuando la distancia es mayor que 295 yardas, las pelotas de MaxFlight pueden ser rechazadas por la USGA por exceder los estándares respecto de distancia de vuelo y carrera.

El programa de control de calidad de MaxFlight consiste en tomar muestras periódicas de 50 pelotas de golf y vigilar el proceso de fabricación. Con cada muestra se realiza una prueba de hipótesis para determinar si el proceso se ha desajustado. Para elaborar las hipótesis nula y alternativa, se empieza por suponer que el proceso está funcionando correctamente; es decir, las pelotas de golf que se están produciendo alcanzan una distancia media de 295 yardas. Ésta es la suposición que establece en la hipótesis nula. La hipótesis alternativa es que la distancia media no es 295 yardas. Como el valor hipotético es $\mu_0 = 295$, las hipótesis nula y alternativa en el caso de la prueba de hipótesis de MaxFlight son las siguientes:

$$H_0: \mu = 295$$
$$H_a: \mu \neq 295$$

Si la media muestral $\bar{x}$ es significativamente menor que 295 yardas o significativamente mayor que 295 yardas, se rechazará H_0. En este caso, se tomarán medidas para ajustar el proceso de fabricación. Por otro lado, si $\bar{x}$ no se desvía una cantidad significativa de la media hipotética $\mu_0 = 295$, H_0 no se rechazará y no se tomará medida alguna para ajustar el proceso de fabricación.

El equipo de control de calidad elige $\alpha = 0.05$ como nivel de significancia para esta prueba. Datos de pruebas previas realizadas sabiendo que el proceso está ajustado indican que se puede suponer que la desviación estándar se conoce y que su valor es $\sigma = 12$. Por ende, como el tamaño de la muestra $n = 50$, el error estándar de $\bar{x}$ es

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{50}} = 1.7$$

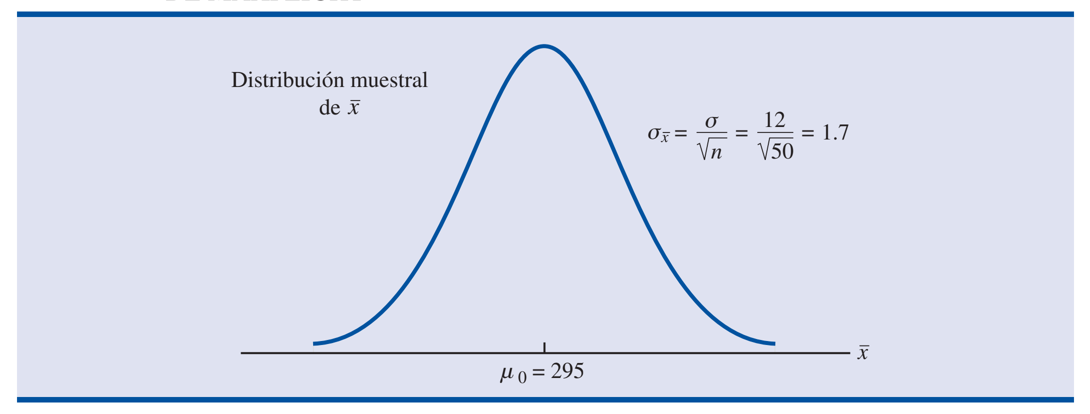
Dado que el tamaño de la muestra es grande, el teorema del límite central (capítulo 7) permite concluir que la distribución muestral de $\bar{x}$ puede aproximarse mediante una distribución normal. En la figura 9.4 se muestra la distribución muestral de x para la prueba de hipótesis de Max Flight con una media poblacional hipotética de $\mu_0 = 295$.



GolfTest

Suponga que se toma una muestra de 50 pelotas de golf y que la media muestral es $\bar{x} = 297.6$ yardas. Esta media muestral favorece la conclusión de que la media poblacional es mayor que 295 yardas. ¿Este valor de $\bar{x}$ es suficientemente mayor que 295 para hacer que se rechace H_0 a un nivel de significancia de 0.05? En la sección anterior se describieron dos métodos que pueden ser usados para responder esta pregunta: el método del valor-p y el método del valor crítico.

FIGURA 9.4 DISTRIBUCIÓN MUESTRAL DE $\bar{x}$ N LA PRUEBA DE HIPÓTESIS DE MAXFLIGHT



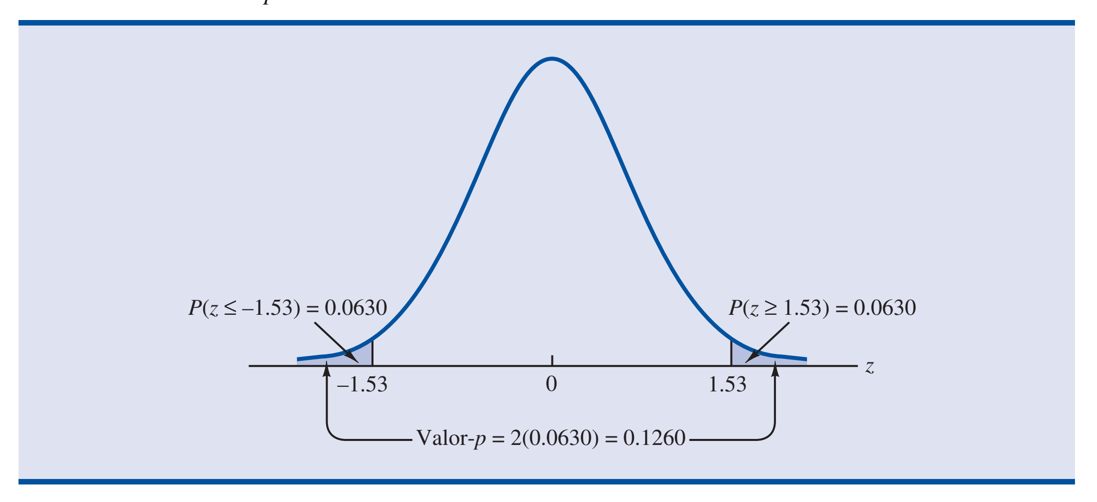
Método del valor-*p* Recuerde que el valor-*p* es la probabilidad que se usa para determinar si se rechaza la hipótesis nula. En una prueba de dos colas, los valores del estadístico de prueba en *ambas* colas, proporcionan evidencias contra la hipótesis nula. En una prueba de dos colas, el valor-*p* es la probabilidad de obtener un valor para el estadístico de prueba *tan improbable o más improbable que* el obtenido con la muestra. A continuación se verá cómo se calcula el valor-*p* en la prueba de hipótesis de MaxFlight.

Primero se calcula el valor del estadístico de prueba. En el caso en que se conoce σ, el estadístico de prueba z es la variable aleatoria normal estándar. Empleando la ecuación (9.1) con $\bar{x} = 297.6$, el valor del estadístico de prueba es

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{297.6 - 295}{12/\sqrt{50}} = 1.53$$

Ahora, para calcular el valor-*p* hay que hallar la probabilidad de obtener, para el estadístico de prueba, un valor *por lo menos tan improbable como* $z = 1.53$. Es claro que los valores $z \geq 1.53$ son *por lo menos tan improbables.* Pero como ésta es una prueba de dos colas, los valores $z \leq -1.53$ también son *por lo menos tan improbables como* el valor del estadístico de prueba obtenido con la muestra. En la figura 9.5 se ve que el valor-*p* para dos colas está dado, en este caso, por $P(z \leq -1.53) + P(z \geq 1.53)$.

FIGURA 9.5 VALOR-*p* EN LA PRUEBA DE HIPÓTESIS DE MAXFLIGHT



Como la curva normal es simétrica, se calcula la probabilidad hallando el área bajo la curva normal a la derecha de $z = 1.53$ y duplicándola. La tabla de la distribución normal estándar indica que el área a la izquierda de $z = 1.53$ es 0.9370. Entonces el área bajo la curva normal a la derecha de $z = 1.53$ es $1.0000 - 0.9370 = 0.0630$. Duplicando esta cantidad, se encuentra que el valor-*p* en la prueba de hipótesis de dos colas de MaxFlight es valor-$p = 2(0.0630) = 0.1260$.

Ahora se compara el valor-*p* con el nivel de significancia para ver si se rechaza la hipótesis nula. Como el nivel de significancia es 0.05, no se rechaza la hipótesis nula porque el valor-*p* es $0.1260 > 0.05$. Como no se rechaza la hipótesis nula, no es necesario tomar medidas para ajustar el proceso de fabricación de MaxFlight.

El cálculo del valor-*p* en una prueba de dos colas puede parecer un poco complicado en comparación con el cálculo del valor-*p* en las pruebas de una cola, pero se simplifica mediante los siguientes pasos.

CÁLCULO DEL VALOR-*p* EN UNA PRUEBA DE DOS COLAS

1. Calcule el valor del estadístico de prueba z.
2. Si el valor del estadístico de prueba está en la cola superior ($z > 0$), encuentre el área bajo la curva normal estándar a la derecha de z. Si el valor del estadístico de prueba está en la cola inferior ($z < 0$), localice el área bajo la curva normal estándar a la izquierda de z.
3. Duplique el área, o probabilidad, en la cola, obtenida en el paso dos y obtenga el valor-*p*.

Método del valor crítico. Antes de dejar esta sección, se verá la forma de comparar el valor del estadístico de prueba z con un valor crítico para tomar la decisión en una prueba de dos colas. En la figura 9.6 se muestra que los valores críticos en esta prueba se encuentran tanto en la cola superior como en la cola inferior de la distribución normal estándar. Si el nivel de significancia es $\alpha = 0.05$, en cada cola, el área más allá del valor crítico es $\alpha/2 = 0.05/2 = 0.025$. En la tabla de probabilidad normal estándar se encuentra que los valores críticos para el estadístico de prueba son $-z_{0.025} = -1.96$ y $z_{0.025} = 1.96$. Entonces, usando el método del valor crítico, la regla de rechazo para dos colas es

$$\text{Rechazar } H_0 \text{ si } z \leq -1.96 \text{ o } z \geq 1.96$$

Como en el estudio de MaxFlight el valor del estadístico de prueba es $z = 1.53$, la evidencia estadística no permitirá rechazar la hipótesis nula a un nivel de significancia de 0.05.

FIGURA 9.6 VALORES CRÍTICOS EN LA PRUEBA DE HIPÓTESIS DE MAXFLIGHT



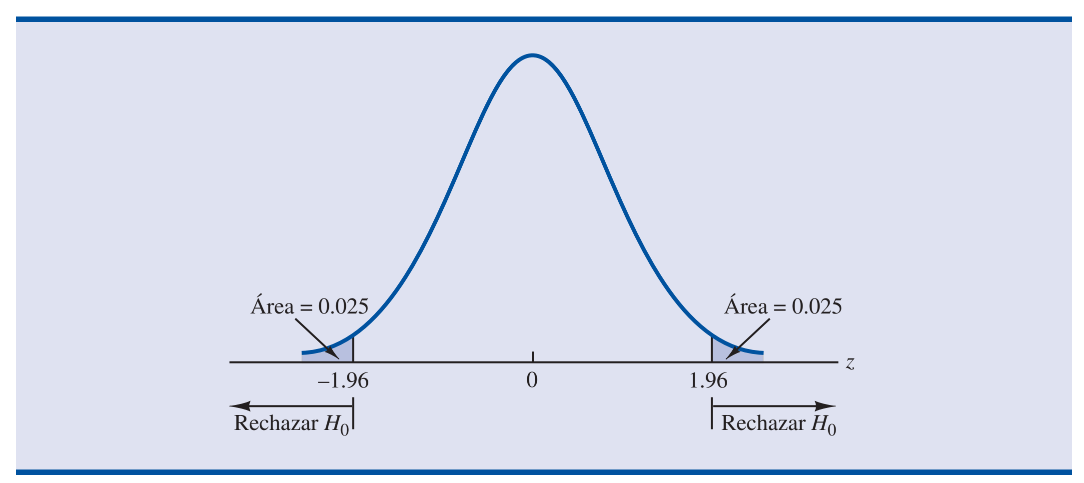


TABLA 9.2 SÍNTESIS DE LAS PRUEBAS DE HIPÓTESIS PARA LA MEDIA POBLACIONAL: CASO σ CONOCIDA

	Prueba de la cola inferior	Prueba de la cola superior	Prueba de dos colas
Hipótesis	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
Estadístico de prueba	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
Regla de rechazo: método del valor-*p*	Rechazar H_0 si valor-$p \leq \alpha$	Rechazar H_0 si valor-$p \leq \alpha$	Rechazar H_0 si valor-$p \leq \alpha$
Regla de rechazo: método del valor crítico	Rechazar H_0 si $z \leq -z_\alpha$	Rechazar H_0 si $z \geq z_\alpha$	Rechazar H_0 si $z \leq -z_{\alpha/2}$ o si $z \geq z_{\alpha/2}$

Resumen y recomendaciones prácticas

Se presentaron ejemplos de una prueba de la cola inferior y de una prueba de dos colas para la media poblacional. Con base en estos ejemplos es posible resumir ahora, como se muestra en la tabla 9.2, los procedimientos para prueba de hipótesis para la media poblacional en el caso de σ conocida.

Los pasos en las pruebas de hipótesis seguidos en los dos ejemplos presentados en esta sección son comunes a toda prueba de hipótesis.

PASOS EN LAS PRUEBAS DE HIPÓTESIS

Paso 1. Dar la hipótesis nula y la hipótesis alternativa.
Paso 2. Especificar el nivel de significancia.
Paso 3. Recabar los datos muestrales y calcular el valor del estadístico de prueba.

Método del valor-p

Paso 4. Emplear el valor del estadístico de prueba para calcular el valor-*p*.
Paso 5. Rechazar H_0 si el valor-$p \leq \alpha$.

Método del valor crítico

Paso 4. Emplear el nivel de significancia para determinar el valor crítico y la regla de rechazo.
Paso 5. Emplear el valor del estadístico de prueba y la regla de rechazo para determinar si se rechaza H_0.

La recomendación práctica acerca del tamaño de la muestra para pruebas de hipótesis es semejante a la recomendación práctica dada en el capítulo 8 respecto del tamaño de la muestra para la estimación por intervalo. En la mayor parte de las aplicaciones, para el procedimiento de prueba de hipótesis visto en esta sección, un tamaño de muestra $n \geq 30$ es adecuado. En los casos en los que el tamaño de la muestra sea menor que 30, la distribución de la población de la que se toma la muestra se vuelve una consideración importante. Si la población tiene una distribución normal, el procedimiento de prueba de hipótesis descrito es exacto y puede usarse con cualquier tamaño de muestra. Si la población no tiene una distribución normal, pero es por lo menos aproximadamente simétrica, con tamaños de muestra hasta de 15 pueden esperarse resultados aceptables.

Relación entre estimación por intervalo y prueba de hipótesis

En el capítulo 8 se mostró la forma de obtener una estimación de la media poblacional mediante un intervalo de confianza. En el caso en que σ es conocida, la estimación de la media poblacional mediante un intervalo de $(1-\alpha)\%$ de confianza está dada por

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

En este capítulo se mostró que una prueba de hipótesis de dos colas para la media poblacional tiene la forma:

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

donde μ_0 es el valor hipotético de la media poblacional.

Suponga que se sigue el procedimiento descrito en el capítulo 8 para construir un intervalo de $(1-\alpha)\%$ de confianza para la media poblacional. Se sabe que $(1-\alpha)\%$ de los intervalos de confianza generados contendrán a la media poblacional y $\alpha\%$ de los intervalos generados no contendrán a la media poblacional. En consecuencia, si se rechaza H_0 siempre que el intervalo de confianza no contenga a μ_0, la probabilidad de rechazar la hipótesis nula cuando sea verdadera $(\mu = \mu_0)$ será α. Recuerde que el nivel de significancia es la probabilidad de rechazar la hipótesis nula cuando es verdadera. Entonces, construir un intervalo de $(1-\alpha)\%$ de confianza y rechazar H_0 siempre que el intervalo no contenga μ_0 es equivalente a realizar una prueba de hipótesis de dos colas con α como nivel de significancia. El procedimiento para usar un intervalo de confianza para una prueba de hipótesis de dos colas se resume como se indica a continuación.

MÉTODO DEL INTERVALO DE CONFIANZA PARA PROBAR UNA HIPÓTESIS DE LA FORMA

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

1. Seleccionar de la población una muestra aleatoria simple y emplear el valor de la media muestral $\bar{x}$ para obtener un intervalo de confianza para la media poblacional μ.

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

2. Si el intervalo de confianza contiene el valor hipotético μ_0, no se rechaza H_0. Si no es así, se rechaza H_0.

En una prueba de hipótesis de dos colas, la hipótesis nula se rechaza si el intervalo de confianza no contiene a μ_0.

El uso del método del intervalo de confianza para realizar una prueba de hipótesis se ilustrará empleando el ejemplo de MaxFlight. La prueba de hipótesis de MaxFlight tiene la forma siguiente:

$$H_0: \mu = 295$$
$$H_a: \mu \neq 295$$

Para probar esta hipótesis con un nivel de confianza $\alpha = 0.05$, se tomó una muestra de 50 pelotas de golf y se halló una distancia media muestral de $\bar{x} = 297.6$ yardas. Recuerde que la desvia-

ción estándar poblacional es $\sigma = 12$. Al aplicar estos resultados a $z_{0.025} = 1.96$, se obtiene que el intervalo de 95% de confianza para estimar la media poblacional es

$$\bar{x} \pm z_{0.025}\frac{\sigma}{\sqrt{n}}$$

$$297.6 \pm 1.96\frac{12}{\sqrt{50}}$$

$$297.6 \pm 3.3$$

o

$$294.3 \text{ a } 300.9$$

Este hallazgo permite al gerente de control de calidad concluir que con 95% de confianza la distancia media para la población de pelotas de golf está entre 294.3 y 300.9 yardas. Como el valor hipotético de la media poblacional es $\mu_0 = 295$, está en dicho intervalo, la conclusión de la prueba de hipótesis es que no se puede rechazar la hipótesis nula, H_0: $\mu = 295$.

Preste atención a que estos análisis y ejemplo pertenecen a pruebas de hipótesis de dos colas para la media poblacional. Sin embargo, la misma relación entre intervalo de confianza y prueba de hipótesis de dos colas existe para otros parámetros poblacionales. Esta relación también se extiende a pruebas de hipótesis de una cola para parámetros poblacionales, para lo que se pide elaborar intervalos de confianza unilaterales que son muy poco usados en la práctica.

NOTAS Y COMENTARIOS

Se mostró cómo usar el valor-*p*. Entre menor sea el valor-*p*, mayor es la evidencia en contra de H_0 y a favor de H_a. Aquí están algunos lineamientos que los profesionistas de la estadística recomiendan para interpretar valores-*p* pequeños.

- Menor que 0.01: Evidencia terminante para concluir que H_a es verdadera.
- Entre 0.01 y 0.05: Fuerte evidencia para concluir que H_a es verdadera.
- Entre 0.05 y 0.10: Evidencia débil para concluir que H_a es verdadera.
- Mayor que 0.10: Evidencia insuficiente para concluir que H_a es verdadera.

Ejercicios

Nota para los estudiantes: en algunos de los ejercicios que siguen se pide usar el método del valor-*p* y en otros el método del valor crítico. Ambos métodos llevarán a la misma conclusión en una prueba de hipótesis. Se presentan ejercicios con ambos métodos para que el estudiante adquiera práctica en el uso de éstos. En las secciones y capítulos posteriores, se preferirá usar el método del valor-*p*, pero el estudiante puede elegir el que prefiera.

Métodos

9. Considere la prueba de hipótesis siguiente:

$$H_0: \mu \geq 20$$
$$H_a: \mu < 20$$

En una muestra de 50, la media muestral fue 19.4. La desviación estándar poblacional es 2.

a. Calcule el valor del estadístico de prueba.
b. ¿Cuál es el valor-p?
c. Use $\alpha = 0.05$, ¿cuál es su conclusión?
d. ¿Cuál es la regla de rechazo si se usa el método del valor crítico? ¿Cuál es su conclusión?



10. Considere la prueba de hipótesis siguiente:

$$H_0: \mu \leq 25$$
$$H_a: \mu > 25$$

En una muestra de 40, la media muestral fue 26.4. La desviación estándar poblacional es 6.

a. Calcule el valor del estadístico de prueba.
b. ¿Cuál es el valor-p?
c. Use $\alpha = 0.01$, ¿cuál es su conclusión?
d. ¿Cuál es la regla de rechazo si se usa el método del valor crítico? ¿Cuál es su conclusión?



11. Considere la prueba de hipótesis siguiente:

$$H_0: \mu = 15$$
$$H_a: \mu \neq 15$$

En una muestra de 50, la media muestral fue 14.15. La desviación estándar poblacional es 3.

a. Calcule el valor del estadístico de prueba.
b. ¿Cuál es el valor-p?
c. Use $\alpha = 0.05$, ¿cuál es su conclusión?
d. ¿Cuál es la regla de rechazo si se usa el método del valor crítico? ¿Cuál es su conclusión?

12. Considere la prueba de hipótesis siguiente:

$$H_0: \mu \geq 80$$
$$H_a: \mu < 80$$

Se usó una muestra de 100, la desviación estándar poblacional es 12. Calcule el valor-p y dé su conclusión para cada uno de los resultados muestrales siguientes. Use $a = 0.01$.

a. $\bar{x} = 78.5$
b. $\bar{x} = 77$
c. $\bar{x} = 75.5$
d. $\bar{x} = 81$

13. Considere la prueba de hipótesis siguiente:

$$H_0: \mu \leq 50$$
$$H_a: \mu > 50$$

Se usó una muestra de 60, la desviación estándar poblacional es 8. Use el valor crítico y dé sus conclusiones para cada uno de los resultados muestrales siguientes. Use $\alpha = 0.05$.

a. $\bar{x} = 52.5$
b. $\bar{x} = 51$
c. $\bar{x} = 51.8$

14. Considere la prueba de hipótesis siguiente:

$$H_0: \mu = 22$$
$$H_a: \mu \neq 22$$

Se usó una muestra de 75, la desviación estándar poblacional es 10. Calcule el valor-p para cada uno de los resultados muestrales siguientes. Use $\alpha = 0.01$.

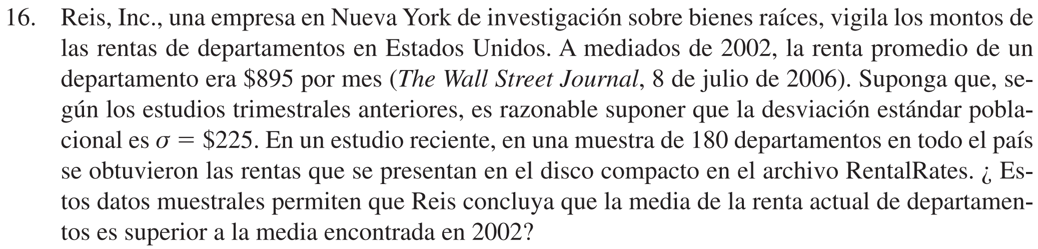
a. $\bar{x} = 23$
b. $\bar{x} = 25.1$
c. $\bar{x} = 20$

Aplicaciones



15. Las declaraciones de impuestos presentadas antes del 31 de marzo obtienen un reembolso que en promedio es de \$1056. Considere la población de los declarantes de "última hora" que presentan su declaración los últimos cinco días del periodo para este trámite (normalmente del 10 al 15 de abril).
    a. Un investigador sugiere que la razón por la que estos declarantes esperan hasta los últimos días se debe a que en promedio obtienen un reembolso menor que los que declaran antes del 31 de marzo. Dé las hipótesis apropiadas de manera que el rechazo de H_0 favorezca la sugerencia de este investigador.
    b. En una muestra de 400 personas que presentaron su declaración entre el 10 y el 15 de abril, la media de los reembolsos fue \$910. Por experiencia se sabe que es posible considerar que la desviación estándar poblacional es $\sigma = \$1600$. ¿Cuál es el valor-$p$?
    c. Con $\alpha = 0.05$, ¿cuál es su conclusión?
    d. Repita la prueba de hipótesis anterior usando el método del valor crítico.





16. Reis, Inc., una empresa en Nueva York de investigación sobre bienes raíces, vigila los montos de las rentas de departamentos en Estados Unidos. A mediados de 2002, la renta promedio de un departamento era \$895 por mes (*The Wall Street Journal*, 8 de julio de 2006). Suponga que, según los estudios trimestrales anteriores, es razonable suponer que la desviación estándar poblacional es $\sigma = \$225$. En un estudio reciente, en una muestra de 180 departamentos en todo el país se obtuvieron las rentas que se presentan en el disco compacto en el archivo RentalRates. ¿ Estos datos muestrales permiten que Reis concluya que la media de la renta actual de departamentos es superior a la media encontrada en 2002?
    a. Dé las hipótesis nula y alternativa.
    b. ¿Cuál es el valor-p?
    c. Con $\alpha = 0.01$, ¿cuál es su conclusión?
    d. ¿Qué le recomendaría a Reis hacer ahora?

17. Las empresas de seguridad de Wall Street pagaron en 2005 gratificaciones de fin de año de \$125 500 por empleado (*Fortune*, 6 de febrero de 2006). Suponga que se desea tomar una muestra de los empleados de la empresa de seguridad Jones & Ryan para ver si la media de la gratificación de fin de año es diferente de la media reportada para la población.
    a. Dé las hipótesis nula y alternativa que usaría para probar si las gratificaciones de fin de año de Jones & Ryan difieren de la media poblacional.
    b. Admita que en una muestra de 40 empleados de Jones & Ryan la media muestral de las gratificaciones de fin de año es \$118 000. Suponga que la desviación estándar poblacional es $\sigma = \$30\,000$ y calcule el valor-p.
    c. Con $\alpha = 0.05$ como nivel de significancia, ¿cuál es su conclusión?
    d. Repita esta prueba de hipótesis usando el método del valor crítico.

18. La rentabilidad anual promedio de los fondos mutualistas U.S. Diversified Equity de 1999 a 2003 fue 4.1% (*BusinessWeek*, 26 de enero de 2004). Un investigador desea realizar una prueba de hipótesis para ver si los rendimientos de determinados fondos de crecimiento (mid-cap growth funds) difieren de manera significativa del promedio de los fondos U.S. Diversified Equity.
    a. Dé las hipótesis que se pueden usar para determinar si la rentabilidad anual media de estos fondos de crecimiento difiere de la media de los fondos U.S. Diversified Equity.
    b. En una muestra de 40 fondos de crecimiento el rendimiento medio fue $\bar{x} = 3.4\%$. Suponga que por estudios anteriores se sabe que la desviación estándar poblacional de estos fondos de crecimiento es $\sigma = 2\%$. Use los resultados muestrales para calcular el estadístico de prueba y el valor-p para la prueba de hipótesis.

c. Con $\alpha = 0.05$, ¿cuál es su conclusión?

19. En 2001, el U.S. Department Labor informó que el salario por hora promedio para los trabajadores de la producción en Estados Unidos era \$14.32 por hora (*The World Almanac 2003*). En 2003, en una muestra de 75 trabajadores de la producción, la media muestral fue \$14.68 por hora. Si la desviación estándar poblacional es $\sigma = \$1.45$, ¿se puede concluir que ha habido un aumento en la media del salario por hora? Use $\alpha = 0.05$.

20. En Estados Unidos un hogar paga en promedio \$32.79 mensuales por el servicio de Internet (CNBC, 18 de enero de 2006). En una muestra de 50 hogares de un estado del sur la media muestral fue \$30.63. Use la desviación estándar poblacional, $\sigma = \$5.60$.
    a. Formule las hipótesis para una prueba en la que se quiere determinar si los datos muestrales favorecen la conclusión de que la cantidad media pagada por el servicio de Internet, en este estado del sur, es menor a la media de todo el país, que es \$32.79.
    b. ¿Cuál es el valor del estadístico de prueba?
    c. ¿Cuál es el valor-*p*?

d. Con $\alpha = 0.01$, ¿cuál es su conclusión?



Fowle

21. Fowle Marketing Research, Inc. tasa la cantidad que cobra a sus clientes en la suposición de que una encuesta por teléfono se realiza en un promedio de 15 minutos o menos. Si se necesita más tiempo en promedio, se cobra una cantidad adicional. Las duraciones de las encuestas en una muestra de 35 encuestas se presentan en el archivo Fowle del disco compacto. Por estudios anteriores se puede considerar que la desviación estándar poblacional es conocida y que es $\sigma = 4$ minutos. ¿Se justifica el cobro de la cantidad adicional?
    d. Formule las hipótesis nula y alternativa para esta aplicación.
    b. Calcule el valor del estadístico de prueba.
    c. ¿Cuál es el valor-*p*?
    d. Con $\alpha = 0.01$, ¿cuál es su conclusión?

22. CNN y ActMedia presentaron un canal de televisión dirigido a las personas que esperan en las colas de los supermercados. En este canal se presentaban noticias, reportajes cortos y publicidad. La duración de la programación estaba basada en la suposición de que la media poblacional del tiempo que los clientes esperan en la cola de la caja era 8 minutos. Se tomará una muestra para verificar si el tiempo medio de espera es realmente 8 minutos.
    a. Formule las hipótesis para esta aplicación.
    b. En una muestra de 120 clientes la media muestral fue 8.5 minutos. Suponga que la desviación estándar poblacional es $\sigma = 3.2$ minutos. ¿Cuál es el valor-*p*?
    c. Con $\alpha = 0.05$, ¿cuál es su conclusión?
    d. Calcule un intervalo de 95% de confianza para la media poblacional. ¿Favorece este intervalo su conclusión?



Media poblacional: σ desconocida

En esta sección se describe cómo realizar pruebas de hipótesis para la media poblacional en el caso de σ desconocida. Como σ desconocida corresponde a la situación en que no se tiene una estimación de la desviación estándar poblacional antes de tomar la muestra, la muestra se usa para obtener una estimación tanto de μ como de σ. Por tanto, para realizar una prueba para la media poblacional en el caso en que no se conoce σ, la media muestral $\bar{x}$ se usa como estimación de μ y la desviación estándar muestral s se usa como estimación de σ.

Los pasos a seguir para las pruebas de hipótesis en el caso en que no se conoce σ son los mismos que en el caso en que se conoce σ, visto en la sección 9.3. Pero como no se conoce σ, los cálculos del estadístico de prueba y del valor-*p* son ligeramente diferentes. Recuerde que en el caso de σ conocida, la distribución muestral del estadístico de prueba tiene distribución normal estándar. Sin embargo, en el caso de σ desconocida la distribución muestral del estadístico de prueba sigue la distribución *t*; tiene ligeramente más variabilidad debido a que la muestra se usa para obtener estimaciones tanto de μ como de σ.

En la sección 8.2 se vio que una estimación por intervalo de la media poblacional en el caso de σ desconocida se basa en una distribución de probabilidad llamada distribución t. Las pruebas de hipótesis para la media poblacional en el caso en que no se conoce σ también se basan en la distribución t. En el caso de σ desconocida el estadístico de prueba tiene distribución t con $n - 1$ grados de libertad.

ESTADÍSTICO DE PRUEBA EN LAS PRUEBAS DE HIPÓTESIS PARA LA MEDIA POBLACIONAL: σ DESCONOCIDA

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{9.2}$$

En el capítulo 8 se dijo que la distribución t se basa en la suposición de que la población de la que se toma la muestra tenga distribución normal. Sin embargo, las investigaciones demuestran que esta suposición no es muy fuerte si el tamaño de la muestra es suficientemente grande. Al final de esta sección se proporciona una recomendación práctica acerca de la distribución de la población y del tamaño de la muestra.

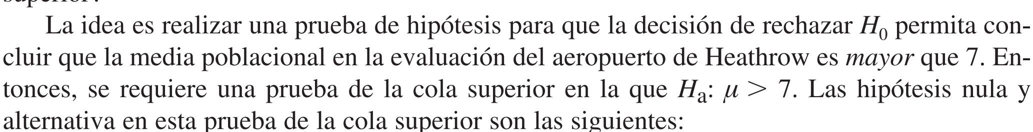
Prueba de una cola

A continuación se considera un ejemplo de prueba de una cola para la media poblacional en el caso de σ desconocida. Una revista de viajes de negocios desea clasificar los aeropuertos internacionales de acuerdo con una evaluación hecha por la población de viajeros de negocios. Se usa una escala de evaluación que va desde un mínimo de 0 hasta un máximo de 10, y aquellos aeropuertos que obtengan una media mayor que 7 serán considerados como aeropuertos de servicio superior. Para obtener los datos de evaluación, el personal de la revista entrevista una muestra de 60 viajeros de negocios de cada aeropuerto. En la muestra tomada en el aeropuerto Heathrow de Londres la media muestral es $\bar{x} = 7.25$ y la desviación estándar muestral es $s = 1.052$. De acuerdo con estos datos muestrales, ¿deberá ser designado Heathrow como un aeropuerto de servicio superior?



AirRating

La idea es realizar una prueba de hipótesis para que la decisión de rechazar H_0 permita concluir que la media poblacional en la evaluación del aeropuerto de Heathrow es *mayor* que 7. Entonces, se requiere una prueba de la cola superior en la que H_a: $\mu > 7$. Las hipótesis nula y alternativa en esta prueba de la cola superior son las siguientes:



$$H_0: \mu \leq 7$$
$$H_a: \mu > 7$$

En esta prueba se usa como nivel de significancia $\alpha = 0.05$.

Al aplicar la ecuación (9.2) con $\bar{x} = 7.25$, $\mu_0 = 7$, $s = 1.052$, y $n = 60$, el valor del estadístico de prueba es

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7.25 - 7}{1.052/\sqrt{60}} = 1.84$$

La distribución muestral de t tiene $n - 1 = 60 - 1 = 59$ grados de libertad. Como es una prueba de la cola superior, el valor-p es el área bajo la curva de la distribución t a la derecha de $t = 1.84$.

Las tablas de la distribución t proporcionada en la mayor parte de los libros de texto no son suficientemente detalladas para determinar el valor-p exacto. Por ejemplo, en la tabla 2 del apéndice B, la distribución t con 59 grados de libertad proporciona la información siguiente.

Área en la cola superior	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
Valor-*t* (59 df)	0.848	1.296	1.671	2.001	2.391	2.662

$t = 1.84$

Como se ve, $t = 1.84$ está entre 1.671 y 2.001. Aunque esta tabla no proporciona el valor exacto de t, los valores que se dan en el renglón "Área en la cola superior" indican que el valor-p debe ser menor que 0.05 y mayor que 0.025. Como el nivel de significancia es 0.05, esto es todo lo que se necesita saber para rechazar la hipótesis nula y concluir que Heathrow debe ser considerado como aeropuerto de servicio superior.

El apéndice F indica cómo calcular los valores-p usando Excel o Minitab.

Dado que es engorroso usar una tabla t para calcular los valores-p y puesto que sólo se pueden obtener valores-p aproximados, se mostrará cómo calcular valores-p exactos usando Excel o Minitab. Estas instrucciones se encuentran al final del libro, en el apéndice F. Usando Excel o Minitab con $t = 1.84$ el valor-p que se obtiene para la cola superior es 0.0354. Como 0.0354 < 0.05, se rechaza la hipótesis nula y se concluye que Heathrow se debe considerar como aeropuerto de servicio superior.

Prueba de dos colas

Para ilustrar cómo realizar una prueba de dos colas para la media poblacional en el caso de σ desconocida se considerará la situación de la prueba de hipótesis de Holiday Toys. Esta empresa distribuye sus productos a través de más de 1 000 puntos de venta. Al planear su producción para la temporada de invierno siguiente, la empresa debe decidir cuántas unidades de cada producto fabricar, antes de saber cuál será la verdadera demanda en cada punto de venta. Para la temporada venidera, el gerente de marketing espera que la demanda de su nuevo juguete sea en promedio 40 unidades por punto de venta. Antes de tomar la decisión final de producción, con base en dicha estimación, la empresa decide hacer una encuesta en una muestra de 25 puntos de venta con objeto de obtener más información acerca de la demanda del nuevo producto. A cada uno de estos puntos de venta se le proporciona información sobre las características del nuevo juguete e información sobre el costo y el precio de venta sugerido. Después se le pide a cada punto de venta que anticipe la cantidad que pedirá.

Los datos muestrales se usan para realizar la siguiente prueba de hipótesis:

$$H_0: \mu = 40$$
$$H_a: \mu \neq 40$$

Si H_0 no se puede rechazar, Holiday continuará con la producción planeada de acuerdo con la estimación del director de marketing de que la media poblacional de la cantidad pedida por punto de venta será $\mu = 40$ unidades. Pero, si H_0 se rechaza, Holiday reevaluará de inmediato su plan de producción de este juguete. Se usa una prueba de dos colas porque Holiday quiere reevaluar su plan de producción si la media poblacional de la cantidad pedida por punto de venta es menor o mayor a la prevista. Como no se cuenta con datos históricos (se trata de un producto nuevo), la media poblacional μ y la desviación estándar poblacional deben estimarse usando los valores $\bar{x}$ y s que se obtengan con los datos muestrales.



En la muestra de 25 puntos de venta la media que se obtiene es $\bar{x} = 37.4$ y la desviación estándar es $s = 11.79$ unidades. Antes de usar la distribución t, el analista elabora un histograma con los datos muestrales con objeto de ver cuál es la forma de la distribución de la población. El histograma de los datos muestrales no muestra evidencias de sesgo ni de valores atípicos, de ma-

nera que el analista concluye que es adecuado usar la distribución t con $n - 1 = 24$ grados de libertad. Usando la ecuación (9.2) con $\bar{x} = 37.4$, $\mu_0 = 40$, $s = 11.79$ y $n = 25$ el valor que se obtiene para el estadístico de prueba es

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{37.4 - 40}{11.79/\sqrt{25}} = -1.10$$

Como se trata de una prueba de dos colas, el valor-p es el área bajo la curva de la distribución t a la izquierda de $t = -1.10$ multiplicado por dos. En la tabla 2 del apéndice B, el renglón de la distribución t para 24 grados de libertad proporciona la información siguiente

Área en la cola superior	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
***Valor t* (24 df)**	0.857	1.318	1.711	2.064	2.492	2.797

$t = 1.10$

La tabla de la distribución t sólo contiene valores t positivos. Sin embargo, como la distribución t es simétrica, el área bajo la curva a la derecha de $t = 1.10$ es igual al área bajo la curva a la izquierda de $t = -1.10$. Se encuentra que $t = 1.10$ está entre 0.857 y 1.318. En el renglón "Área en la cola superior" se ve que el área en la cola a la derecha de $t = 1.10$ está entre 0.20 y 0.10. Duplicando estas cantidades, el valor-p debe estar entre 0.40 y 0.20. Como el nivel de significancia es $\alpha = 0.05$, se ve que el valor-p es mayor que α. Por tanto, no se puede rechazar H_0. No hay evidencia suficiente para concluir que Holiday deba modificar su plan de producción para la temporada siguiente.

En el apéndice F se indica cómo calcular el valor-p para esta prueba usando Minitab o Excel. El valor-p que se obtiene es 0.2822. Con el nivel de significancia $\alpha = 0.05$, no se puede rechazar H_0 dado que $0.2822 > 0.05$.

Para tomar la decisión en esta prueba de dos colas también se puede comparar el estadístico de prueba con el valor crítico. Usando $\alpha = 0.05$ y la distribución t con 24 grados de libertad, $-t_{0.025} = -2.064$ y $t_{0.025} = 2.064$ son los valores críticos para la prueba de dos colas. La regla de rechazo usando el estadístico de prueba es

$$\text{Rechazar } H_0 \text{ si } t \leq -2.064 \text{ o si } t \geq 2.064$$

De acuerdo con el estadístico de prueba $t = -1.10$, H_0 no puede rechazarse. Este resultado indica que Holyday puede continuar con su plan de producción para la temporada próxima de acuerdo con la expectativa de $\mu = 40$.

Resumen y recomendación práctica

En la tabla 9.3 se proporciona un resumen de los procedimientos de prueba de hipótesis en los casos de σ desconocida. La diferencia principal entre estos procedimientos para el caso σ conocida es que para calcular el estadístico de prueba se usa s en lugar de σ. A esto se debe que el estadístico de prueba siga la distribución t.

La aplicabilidad de los procedimientos de prueba de hipótesis de esta sección depende de la distribución de la población de donde se toma la muestra y del tamaño de la muestra. Si la población tiene una distribución normal, las pruebas de hipótesis descritas en esta sección dan resultados exactos con cualquier tamaño de muestra. Si la población no está distribuida normalmente, los procedimientos son aproximaciones. De cualquier manera, se encuentra que tamaños de muestra de 30 o mayores proporcionan buenos resultados en la mayor parte de los casos. Si la población es aproximadamente normal, muestras pequeñas (por ejemplo, $n < 15$) pueden proporcionar resultados aceptables. Si la población es muy sesgada o si contiene observaciones atípicas, se recomiendan tamaños de muestra de alrededor de 50.

TABLA 9.3 SÍNTESIS DE LAS PRUEBAS DE HIPÓTESIS PARA LA MEDIA POBLACIONAL: CASO σ DESCONOCIDA

	Prueba de la cola inferior	**Prueba de la cola superior**	**Prueba de dos colas**
Hipótesis	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
Estadístico de prueba	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Regla de rechazo: *método del valor-p*	Rechazar H_0 si valor-$p \leq \alpha$	Rechazar H_0 si valor-$p \leq \alpha$	Rechazar H_0 si valor-$p \leq \alpha$
Regla de rechazo: método del valor crítico	Rechazar H_0 si $t \leq -t_\alpha$	Rechazar H_0 si $t \geq t_\alpha$	Rechazar H_0 si $t \leq -t_{\alpha/2}$ o si $t \geq t_{\alpha/2}$

Ejercicios

Métodos

23. Considere la prueba de hipótesis siguiente:

$$H_0: \mu \leq 12$$
$$H_a: \mu > 12$$

En una muestra de 25, la media muestral fue $\bar{x} = 14$ y la desviación estándar muestral fue $s = 4.32$.

a. Calcule el valor del estadístico de prueba.
b. Use la tabla de la distribución t (tabla 2 del apéndice B) para calcular un intervalo para el valor-p.
c. Con $\alpha = 0.05$, ¿cuál es su conclusión?
d. ¿Cuál es la regla de rechazo usando el valor crítico? ¿Cuál es su conclusión?



24. Considere la prueba de hipótesis siguiente:

$$H_0: \mu = 18$$
$$H_a: \mu \neq 18$$

En una muestra de 48, la media muestral fue $\bar{x} = 17$ y la desviación estándar muestral fue $s = 4.5$.

a. Calcule el valor del estadístico de prueba.
b. Use la tabla de la distribución t (tabla 2 del apéndice B) para calcular un intervalo para el valor-p.
c. Con $\alpha = 0.05$, ¿cuál es su conclusión?
d. ¿Cuál es la regla de rechazo usando el valor crítico? ¿Cuál es su conclusión?

25. Considere la prueba de hipótesis siguiente:

$$H_0: \mu \geq 45$$
$$H_a: \mu < 45$$

Se usa una muestra de 36. Identifique el valor-p y establezca su conclusión para cada uno de los siguientes resultados muestrales. Use $\alpha = 0.01$.

a. $\bar{x} = 44$ y $s = 5.2$
b. $\bar{x} =43$ y $s = 4.6$
c. $\bar{x} = 46$ y $s = 5.0$

26. Considere la prueba de hipótesis siguiente:

$$H_0: \mu = 100$$
$$H_a: \mu \neq 100$$

Se usa una muestra de 65. Identifique el valor-p y establezca su conclusión para cada uno de los siguientes resultados muestrales. Use $\alpha = 0.05$.

a. $\bar{x} = 103$ y $s = 11.5$
b. $\bar{x} = 96.5$ y $s = 11.0$
c. $\bar{x} = 102$ y $s = 10.5$

Aplicaciones



27. La Employment and Training Administration informó que la prestación media del seguro de desempleo es \$238/semana (*The World Almanac*, 2003). Un investigador del estado de Virginia anticipó que datos muestrales indicarán que la prestación media semanal del seguro de desempleo en el estado de Virginia es menor que la media de todo el país.
   a. Dé las hipótesis adecuadas de manera que el rechazo de H_0 favorezca la afirmación del investigador.
   b. En una muestra de 100 individuos la media muestral encontrada fue \$231 y la desviación estándar muestral fue \$80. ¿Cuál es el valor-p?
   c. Si $\alpha = 0.05$, ¿cuál es su conclusión?
   d. Repita la prueba de hipótesis anterior usando el método del valor crítico.

28. La Asociación Nacional de Ligas de Béisbol Profesional de Estados Unidos, informó que en la temporada de 2001 la asistencia a 176 juegos de béisbol de liga menor alcanzó un máximo sin precedentes (*New York Times*, 28 de julio de 2002). La asistencia promedio a un juego de béisbol fue de 3530 personas por juego. A la mitad de la temporada de 2002, el presidente de la asociación solicitó un informe de asistencia con la esperanza de que superara a la asistencia del 2001.
   a. Formule las hipótesis que se usarán para determinar si la asistencia media por juego en el 2002 excedieron a las del año anterior.
   b. Suponga que en una muestra de 92 juegos de béisbol de la liga menor jugados en la primera mitad de la temporada de 2002, la asistencia media es de 3740 personas por juego y la desviación estándar 810. ¿Cuál es el valor-p?
   c. Si $\alpha = 0.01$, ¿cuál es su conclusión?



29. El precio de un diamante de un quilate de color H y pureza VS2 de Diamod Source USA es \$5 600 (www.diasource.com, marzo de 2003). Un joyero del medio oeste llama al distrito de los diamantes de Nueva York para ver si el precio medio de los diamantes ahí difiere de \$5600
   a. Formule las hipótesis para determinar si el precio en Nueva York difiere de \$5600.
   b. Los precios en una muestra de 25 contactos en la ciudad de Nueva York fueron los que se presentan en el archivo *Diamonds* del disco compacto. ¿Cuál es el valor-p?
   c. Con $\alpha = 0.05$, ¿es posible rechazar la hipótesis nula? ¿Cuál es su conclusión?
   d. Repita la prueba de hipótesis anterior usando el método del valor crítico.

30. CNN una compañía de AOL Time Warner Inc. Tiene el liderazgo de noticias en la televisión por cable. Nielsen Media Research indica que en 2002 la media de la audiencia de CNN fue de 600 000 espectadores por día. Suponga que en una muestra de 40 días durante la primera mitad de 2003, la cantidad diaria de espectadores haya sido 612 000 espectadores por día y la desviación estándar 65 000 espectadores.
   a. ¿Cuáles son las hipótesis si el director de CNN desea información sobre cualquier cambio en la cantidad de espectadores de la CNN?
   b. ¿Cuál es el valor-p?
   c. Elija su propio nivel de significancia. ¿Cuál es su conclusión?
   d. ¿Qué recomendación le haría al director de CNN en esta aplicación?

31. Raftelis Financial Consulting informa que la media en los recibos trimestrales del agua en Estados Unidos es \$47.50 (*U.S. News & World Report*, 12 de agosto de 2002). Algunos servicios de agua son operados por empresas de servicio público, mientras que otros sistemas de agua son operados por empresas privadas. Un economista indica que la privatización no nivela la competencia y que el poder monopólico dado a las empresas públicas se está transfiriendo ahora a las empresas privadas. El problema es que los usuarios acaban pagando tasas más altas por el agua suministrada por las empresas privadas. El sistema de agua de Atlanta, Georgia, es operado por una empresa privada. En una muestra de 64 usuarios de Atlanta, la cantidad media trimestral pagada por el agua fue \$51 y la desviación estándar fue \$12. Empleando $\alpha = 0.05$ ¿la muestra favorece la conclusión de que esta empresa privada que suministra el agua tiene tasas promedio mayores?



UsedCars

32. De acuerdo con la National Automobile Dealers Association, el precio medio de un automóvil usado es \$10 192. El administrador de una distribuidora de la ciudad de Kansas revisó una muestra de 50 automóviles usados vendidos en esa distribuidora recientemente, con objeto de determinar si la media poblacional de sus precios difería del precio medio en todo el país. Los precios de los 50 automóviles se encuentran en el disco compacto en el archivo denominado *UsedCars*.
    a. Formule las hipótesis para determinar si existe diferencia en el precio medio de los automóviles usados de la distribuidora.
    b. ¿Cuál es el valor-p?
    c. Con $\alpha = 0.05$, ¿cuál es su conclusión?

33. El consumo anual per cápita de leche es 21.6 galones (*Statistical Abstract of the United States: 2006*). Usted cree que en el oeste medio el consumo de leche es mayor y desea fundamentar su opinión. En una muestra de 16 personas de Webster City, pueblo del oeste medio, la media muestral del consumo anual fue de 24.1 galones y la desviación estándar es $s = 4.8$.
    a. Elabore una prueba de hipótesis que se pueda usar para determinar si el consumo medio anual en Webster City es mayor que la media nacional.
    b. Dé una estimación puntual de la diferencia entre el consumo medio anual en Webster City y el consumo medio anual nacional.
    c. Con $\alpha = 0.05$ pruebe si hay una diferencia significativa. ¿Cuál es su conclusión?

34. Joan´s Nursery se especializa en jardines de zonas residenciales, de acuerdo con el diseño del cliente. La estimación del precio de un proyecto se basa en el número de árboles, arbustos, etc., a emplear en el proyecto. Para propósitos de estimación de costos, los administradores consideran que se requieren dos horas de trabajo para plantar un árbol mediano. A continuación se presentan los tiempos (en horas) realmente requeridos en una muestra de 10 árboles plantados el mes pasado.

1.7	1.5	2.6	2.2	2.4	2.3	2.6	3.0	1.4	2.3

    Utilice el nivel de significancia $\alpha = 0.05$, realice una prueba para ver si el tiempo necesario promedio para plantar los árboles difiere de 2 horas.
    a. Establezca las hipótesis nula y alternativa.
    b. Calcule la media muestral.
    c. Calcule la desviación estándar muestral.
    d. ¿Cuál es el valor-p?
    e. ¿Cuál es su conclusión?

9.5 Proporción poblacional

En esta sección se muestra cómo realizar una prueba de hipótesis para la proporción poblacional p. Mediante p_0 se denota la proporción poblacional, las tres formas de una prueba de hipótesis para la proporción poblacional son las siguientes:

$$H_0: p \geq p_0 \qquad H_0: p \leq p_0 \qquad H_0: p = p_0$$
$$H_a: p < p_0 \qquad H_a: p > p_0 \qquad H_a: p \neq p_0$$

La primera forma es una prueba de la cola inferior, la segunda es una prueba de la cola superior y la tercera es una prueba de dos colas.

Las pruebas de hipótesis para la proporción poblacional se basan en la diferencia entre la proporción muestral $\bar{p}$ y la proporción poblacional hipotética p_0. Los métodos para realizar la prueba de hipótesis son semejantes a los usados para las pruebas de hipótesis para la media poblacional. La única diferencia es que para calcular el estadístico de prueba se usa la proporción muestral y su error estándar. Después, para determinar si se rechaza la hipótesis nula se usa el método del valor-p o el método del valor crítico.

Para ver un ejemplo, se considera el caso del campo de golf Pine Creek. En los años anteriores 20% de los jugadores del campo eran mujeres. Para aumentar la proporción de mujeres, se realizó una promoción especial. Un mes después de realizada la promoción, el directivo del campo solicita un estudio estadístico para determinar si la proporción de jugadoras ha aumentado. Como el objetivo es determinar si la proporción de jugadoras ha aumentado, lo apropiado es una prueba de la cola derecha en la que H_a: $p > 0.20$. Las hipótesis nula y alternativa para esta prueba son:

$$H_0: p \leq 0.20$$
$$H_a: p > 0.20$$

Si se puede rechazar H_0 los resultados de la prueba darán sustento estadístico a la conclusión de que la proporción de golfistas aumentó y que la promoción fue efectiva. El directivo del campo especificó que se usara $\alpha = 0.05$ como nivel de significancia para realizar dicha prueba.

El paso siguiente en el procedimiento de prueba de hipótesis es seleccionar una muestra y calcular el valor del estadístico de prueba adecuado. Para demostrar cómo se realiza este paso en la prueba de la cola superior se empieza por calcular el valor del estadístico de prueba en cualquier forma de prueba de hipótesis para la proporción poblacional. La distribución muestral de $\bar{p}$, el estimador puntual del parámetro poblacional p, es la base para desarrollar el estadístico de prueba.

Si la hipótesis nula es verdadera como igualdad, el valor esperado de $\bar{p}$ es igual al valor hipotético p_0; es decir, $E(\bar{p}) = p_0$. El error estándar de $\bar{p}$ está dado por

$$\sigma_{\bar{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

En el capítulo 7 se dijo que si $np \geq 5$ y $n(1 - p) \geq 5$, la distribución muestral de $\bar{p}$ puede aproximarse mediante una distribución normal.* Bajo estas condiciones que normalmente aplican en la práctica, la igualdad

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} \tag{9.3}$$

tiene distribución de probabilidad normal estándar. Con $\sigma_{\bar{p}} = \sqrt{p_0(1 - p_0)/n}$, la variable aleatoria normal estándar z es el estadístico de prueba empleado para realizar las pruebas de hipótesis acerca de la proporción poblacional.

*En la mayor parte de las aplicaciones de pruebas de hipótesis para la proporción poblacional, los tamaños de las muestras son suficientemente grandes para usar la aproximación a la distribución normal. La distribución muestral exacta de $\bar{p}$ es discreta y la probabilidad para cada valor de $\bar{p}$ está dada por la distribución binomial. En consecuencia, las pruebas de hipótesis son un poco más complicadas cuando las muestras son pequeñas y no se puede usar la aproximación a la distribución normal.

ESTADÍSTICO DE PRUEBA EN LAS PRUEBAS DE HIPÓTESIS PARA LA PROPORCIÓN POBLACIONAL

$$z = \frac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} \tag{9.4}$$



Ahora es posible calcular el estadístico de prueba para la prueba de hipótesis del campo de golf Pine Creek. Considere una muestra de 400 jugadores en la que 100 de los jugadores son mujeres. La proporción de mujeres golfistas en la muestra es

$$\bar{p} = \frac{100}{400} = 0.25$$

Al aplicar la ecuación (9.4), el valor del estadístico de prueba es

$$z = \frac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \frac{0.25 - 0.20}{\sqrt{\dfrac{0.20(1 - 0.20)}{400}}} = \frac{0.05}{0.02} = 2.50$$

Como la prueba de hipótesis para el campo de golf es una prueba de la cola superior, el valor-p es la probabilidad de que z sea mayor o igual que $z = 2.50$. En la tabla de probabilidad normal estándar aparece que el área a la izquierda de $z = 2.50$ es 0.9938. Por tanto, el valor-p en la prueba del campo de golf es $1.0000 - 0.9938 = 0.0062$. En la figura 9.7 se muestra el cálculo de este valor-p.

Recuerde que el administrador del campo especificó $\alpha = 0.05$ como nivel de significancia. Un valor-$p = 0.0062 < 0.05$ proporciona evidencia estadística suficiente para rechazar H_0 al nivel de significancia 0.05. Así, la prueba da apoyo estadístico suficiente para la conclusión de que la promoción especial incrementó la proporción de jugadoras en el campo de golf.

La decisión de rechazar o no la hipótesis nula también se toma usando el método del valor crítico. El valor crítico que corresponde a un área de 0.05 en la cola superior de una distribución de probabilidad normal es $z_{0.05} = 1.645$. Entonces, la regla de rechazo usando el método del valor crítico es rechazar H_0 si $z \geq 1.645$. Como $z = 2.50 > 1.645$, se rechaza H_0.

Una vez más, los métodos del valor-p y del valor crítico llevan a la misma conclusión en una prueba de hipótesis, pero el método del valor-p proporciona más información. Para un valor-p =

FIGURA 9.7 CÁLCULO DEL VALOR-p PARA LA PRUEBA DE HIPÓTESIS DEL CAMPO DE GOLF

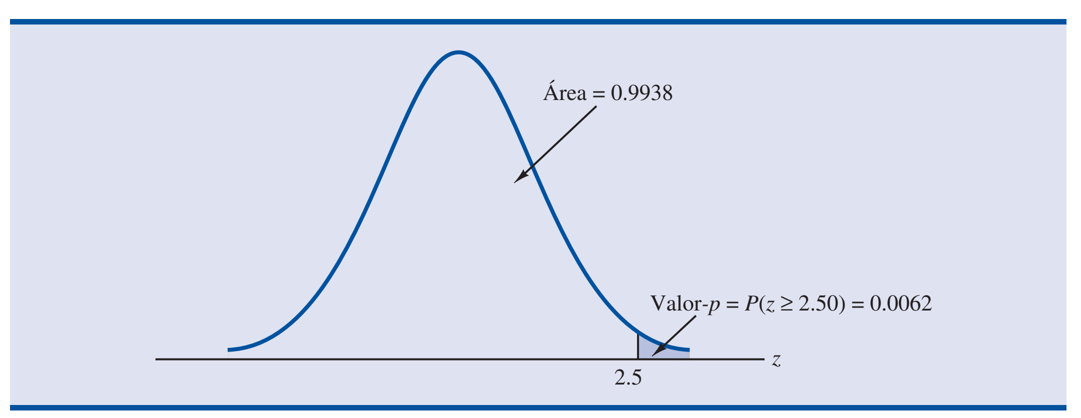


TABLA 9.4 SÍNTESIS DE LAS PRUEBAS DE HIPÓTESIS PARA LA MEDIA POBLACIONAL

	Prueba de la cola inferior	Prueba de la cola superior	Prueba de dos colas
Hipótesis	$H_0: p \geq p_0$ $H_a: p < p_0$	$H_0: p \leq p_0$ $H_a: p > p_0$	$H_0: p = p_0$ $H_a: p \neq p_0$
Estadístico de prueba	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$
Regla de rechazo: método del valor-*p*	Rechazar H_0 si valor-$p \leq \alpha$	Rechazar H_0 si valor-$p \leq \alpha$	Rechazar H_0 si valor-$p \leq \alpha$
Regla de rechazo: método del valor crítico	Rechazar H_0 si $z \leq -z_\alpha$	Rechazar H_0 si $z \geq z_\alpha$	Rechazar H_0 si $z \leq -z_{\alpha/2}$ o si $z \geq z_{\alpha/2}$

0.0062, la hipótesis nula será rechazada para cualquier nivel de significancia mayor o igual que 0.0062.

Resumen

El procedimiento usado en una prueba de hipótesis para la proporción poblacional es semejante al método usado para una prueba de hipótesis para la media poblacional. Aunque sólo se ilustró cómo realizar una prueba de hipótesis de la cola superior para la proporción poblacional, para pruebas de la cola inferior o para pruebas de dos colas se usan procedimientos similares. En la tabla 9.4 se presenta una síntesis de las pruebas de hipótesis para proporción poblacional. Se supone que $np \geq 5$ y $n(1 - p) \geq 5$; con lo cual se puede usar una distribución normal como aproximación a la distribución muestral de $\bar{p}$.

Ejercicios

Métodos

35. Considere la prueba de hipótesis siguiente:

$$H_0: p = 0.20$$
$$H_a: p \neq 0.20$$

En una muestra de 400 se encontró la proporción muestral $\bar{p} = 0.175$.
a. Calcule el valor del estadístico de prueba.
b. ¿Cuál es el valor-p?
c. Con $\alpha = 0.05$, ¿cuál es su conclusión?
d. ¿Cuál es la regla de rechazo usando el valor crítico? ¿Cuál es su conclusión?



36. Considere la prueba de hipótesis siguiente:

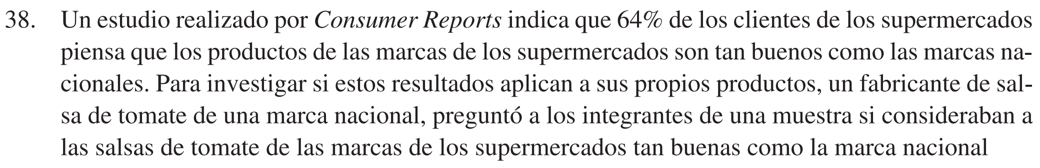
$$H_0: p \geq 0.75$$
$$H_a: p < 0.75$$

Se seleccionó una muestra de 300 elementos. Calcule el valor-p y establezca su conclusión para cada uno de los resultados muestrales siguientes. Use $\alpha = 0.05$.
a. $\bar{p} = 0.68$ c. $\bar{p} = 0.70$
b. $\bar{p} = 0.72$ d. $\bar{p} = 0.77$

Aplicaciones

37. En un estudio se encontró que en 2005, 12.5% de los trabajadores estadounidenses pertenecían a un sindicato (*The Wall Street Journal*, 21 de enero de 2006). El caso es que en 2006 se toma una muestra de 400 trabajadores estadounidenses para ver si el esfuerzo realizado por los sindicatos por organizarse ha hecho que aumente el número de sus miembros.
   a. Formule las hipótesis que puedan ser usadas para determinar si la membresia de los sindicatos ha aumentado en 2006.
   b. Si los resultados muestrales indican que 52 de los trabajadores pertenecen a los sindicatos, ¿cuál es el valor-p de esta prueba de hipótesis?
   c. Con $\alpha = 0.05$, ¿cuál es su conclusión?





38. Un estudio realizado por *Consumer Reports* indica que 64% de los clientes de los supermercados piensa que los productos de las marcas de los supermercados son tan buenos como las marcas nacionales. Para investigar si estos resultados aplican a sus propios productos, un fabricante de salsa de tomate de una marca nacional, preguntó a los integrantes de una muestra si consideraban a las salsas de tomate de las marcas de los supermercados tan buenas como la marca nacional
   a. Formule las hipótesis para determinar si el porcentaje de clientes de los supermercados que considera a las salsas de tomate de las marcas de los supermercados tan buenas como la marca nacional difiere de 64%.
   b. Si en una muestra de 100 clientes 52 opinan que las marcas de los supermercados son tan buenas como las marcas nacionales, ¿cuál es el valor-p?
   c. Con $\alpha = 0.05$, ¿cuál es la conclusión?
   d. ¿Le dará gusto esta conclusión al fabricante de la marca nacional de salsa de tomate? Explique.

39. El National Center for Health Statistics reportó que 70% de los adultos no hacen ejercicio con regularidad. Un investigador decide realizar un estudio para ver si esto difiere de un estado a otro.
   a. Establezca las hipótesis nula y alternativa si la intención del investigador es identificar los estados que difieren de este 70% reportado.
   b. Con $\alpha = 0.05$, cuál es la conclusión en los estados siguientes:

Wisconsin:	252 de 350 adultos no hacen ejercicio con regularidad.
California:	189 de 300 adultos no hacen ejercicio con regularidad.

40. Antes del Super Bowl de 2003, la ABC pronosticó que 22% de la audiencia por televisión expresaría interés por ver uno de sus próximos programas: *8 Simple Rules*, *Are You Hot?* y *Dragnet*. Durante el Super Bowl, la ABC pasó comerciales sobre estos programas de televisión. Al día siguiente del Super Bowl, una empresa de publicidad tomó una muestra de 1 532 espectadores que los vieron, de los cuales 414 afirmaron que verían alguna de las series promovidas por la ABC.
   a. ¿Cuál es la estimación puntual de la proporción de espectadores que después de ver los comerciales sobre los programas de televisión dijeron que los verían?
   b. Con $\alpha = 0.05$, determine si la intención de ver los programas de la ABC aumentó significantemente después de ver los comerciales.
   c. ¿Por qué tales estudios son valiosos para las empresas y los negocios de publicidad?

41. En una conferencia en 2006 un ejecutivo, de una empresa mediadora en el mercado de dinero, dijo a un grupo de analistas que 70% de los inversionistas confían en lograr sus objetivos de inversión. UBS Investors Optimism Survey realizó un estudio, del 2 al 15 de enero, y encontró que 67% de los inversionistas confiaban en lograr sus objetivos de inversión.
   a. Formule las hipótesis para probar la validez de lo dicho por el ejecutivo de la empresa mediadora en el mercado de dinero.

b. Suponga que para este estudio se reunió información de 300 inversionistas. ¿Cuál es el valor-p en esta prueba de hipótesis?
c. Con $\alpha = 0.50$. ¿Debe rechazarse lo que afirma el ejecutivo?

42. De acuerdo con un estudio realizado por el Census Bureau´s Americn Housing Survey, cuando una persona se muda de casa, el factor principal en la elección de su nuevo domicilio es que esté cerca de su trabajo (*USA Today*, 24 de diciembre de 2002). Según datos de 1990 de la Census Bureau, se sabe que 24% de la población de personas que se muda de casa da una “ubicación cercana a su trabajo” como el factor principal en la selección de su nuevo domicilio. Considere que en una muestra de 300 personas que se mudaron de casa en 2003, 93 lo hicieron para estar más cerca de su trabajo. ¿Los datos muestrales respaldan la conclusión de la investigación de que en 2003 hay más personas que buscan un domicilio cercano a su trabajo? ¿Cuál es la estimación puntual de la proporción de personas que se mudaron en 2003 buscando estar más cerca de su trabajo? ¿Cuál es la conclusión de la investigación? Use $\alpha = 0.05$.



Eagle

43. Eagle Outfitters es una cadena de tiendas que se especializa en ropa de invierno y equipo para excursionismo. Esta empresa planea una promoción con envío de cupones de descuento para todos sus clientes con tarjeta de crédito. La promoción será un éxito si más de 10% de los que reciban el cupón lo utilizan. Antes de realizar la promoción a nivel nacional, se envían cupones a los integrantes de una muestra de 100 clientes con tarjeta de crédito.
    a. Dé las hipótesis que pueden ser usadas para probar si la proporción poblacional de aquellos que usarán el cupón es suficiente como para hacer la promoción en todo el país.
    b. El archivo Eagle contiene los datos muestrales. Dé una estimación puntual de la proporción poblacional.
    c. Use $\alpha = 0.05$ y realice la prueba de hipótesis. ¿La empresa debe realizar esta promoción en todo el país?



Drowsy

44. En un artículo anunciado en portada, *BusinessWeek* publicó información acerca de los hábitos de sueño de los estadounidenses (*BusinessWeek*, 26 de enero de 2004). El artículo señalaba que la privación del sueño, ocasiona diversos problemas, entre ellos muertes en las autopistas. Cincuenta y uno por ciento de los conductores admitió manejar sintiéndose somnoliento. Un investigador planteó la hipótesis de que este problema es aún mayor entre los trabajadores de los turnos nocturnos.
    a. Formule las hipótesis que ayuden a determinar si más de 51% de la población de trabajadores de los turnos nocturnos admiten conducir somnolientos.
    b. En una muestra de 500 trabajadores de turnos nocturnos. se identificó a quienes admitían conducir somnolientos ¿Cuál es la proporción muestral? ¿Cuál es el valor-p?
    c. Con $\alpha = 0.1$, ¿cuál es la conclusión?

45. Muchos inversionistas y analistas financieros piensan que el Promedio Industrial Dow Jones (DJIA) es un buen barómetro del mercado de acciones. El 31 de enero de 2006, 9 de las 30 acciones que constituyen el DJIA subieron de precio (*The Wall Street Journal*, 1 de febrero de 2006). A partir de este hecho, afirmó que 30% de las acciones de la Bolsa de Nueva York subirían ese mismo día.
    a. Formule las hipótesis nula y alternativa para probar lo que afirma el analista.
    b. En una muestra de 50 acciones de la bolsa de Nueva York, 24 subieron. Dé la estimación puntual de la proporción poblacional de las acciones que subieron.
    c. Realice una prueba de hipótesis usando $\alpha = 0.01$ como nivel de significancia. ¿Cuál es la conclusión?



Prueba de hipótesis y toma de decisiones

En la sección 9.1 se vieron tres tipos de situaciones en las que se usa una prueba de hipótesis.

1. Para probar una hipótesis de investigación.
2. Para probar la validez de una afirmación.
3. Para tomar una decisión.

En las dos primeras situaciones sólo se toma alguna acción si se rechaza la hipótesis nula H_0, por lo que se concluye que la hipótesis alternativa es verdadera. En la tercera situación —toma de decisiones— es necesario tomar alguna acción tanto si se acepta como si se rechaza la hipótesis nula.

La aplicabilidad de los procedimientos de prueba de hipótesis, considerados hasta ahora, es limitada para la toma de decisiones, porque no se considera apropiado aceptar H_0 y tomar medidas con base en la conclusión de que H_0 es verdadera. La razón para no tomar medidas cuando el resultado de la prueba indica *no rechazar* H_0 es que la decisión de aceptar H_0 expone a quien toma la decisión al riesgo de cometer un error tipo II; a aceptar H_0 cuando es falsa. En los procedimientos de prueba de hipótesis descritos en las secciones anteriores, se controla la probabilidad de cometer un error tipo I al establecer el nivel de significancia para la prueba. Pero la probabilidad de cometer un error tipo II no se controla.

Si es necesario tomar una decisión, quien debe tomar la decisión algunas veces deseará —y en muchos casos tendrá que— actuar tanto en el caso en que la conclusión sea *no rechazar* H_0 como en el caso en que la decisión sea *rechazar* H_0. Una buena ilustración de esta situación es el muestreo de aceptación, un tema que se discutirá con más detalle en el capítulo 20. Por ejemplo, el director de control de calidad tiene que decidir si acepta un pedido de baterías recibido de un proveedor o si lo rechaza por ser de mala calidad. Las especificaciones indican que la vida útil promedio de las baterías debe ser por lo menos 120 horas. Para evaluar si el pedido recibido satisface esta especificación, se selecciona una muestra de 36 baterías y se prueban. De acuerdo con esta muestra se deberá tomar la decisión de aceptar el pedido de baterías o devolverlo por no tener la calidad adecuada. Sea μ el número medio de horas de vida útil que tienen las baterías del envío. Las hipótesis nula y alternativa para la media poblacional serán las que se presentan a continuación.

$$H_0: \mu \geq 120$$
$$H_a: \mu < 120$$

Si se rechaza H_0, se concluye que la hipótesis alternativa es verdadera. Esta conclusión indica que lo adecuado es devolver el pedido al proveedor. Pero si no se rechaza H_0, la persona que toma la decisión deberá determinar qué medidas tomar. Así, sin haber concluido que H_0 es verdadera, sino sólo por no haberla rechazado, la persona que toma la decisión habrá de aceptar el envío y considerarlo de la calidad adecuada.

En tales situaciones, es recomendable que el procedimiento de prueba de hipótesis se amplíe para controlar la probabilidad de cometer un error tipo II. Como se tomará una decisión y alguna medida cuando no se rechace H_0, será útil conocer la probabilidad de cometer un error tipo II. En las secciones 9.7 y 9.8 se explica cómo calcular la probabilidad de cometer un error tipo II y ajustar el tamaño de la muestra para controlar la probabilidad de cometer un error tipo II.



Cálculo de la probabilidad de los errores tipo II

En esta sección se muestra cómo calcular la probabilidad de cometer un error tipo II en una prueba de hipótesis para la media poblacional. Este procedimiento se ilustra usando el ejemplo del muestreo de aceptación descrito en la sección 9.6. Las hipótesis nula y alternativa para el número medio de horas de vida útil de un pedido de baterías son: H_0: $\mu \geq 120$ y H_a: $\mu < 120$. Si se rechaza H_0, la decisión será devolver el pedido al proveedor debido a que la media del número de horas de vida útil es menor que 120 horas, lo especificado. Si no rechaza H_0, la decisión será aceptar el pedido.

Para realizar la prueba de hipótesis se usa como nivel de significancia $a = 0.05$. El estadístico de prueba en el caso σ conocida es

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 120}{\sigma/\sqrt{n}}$$

Con base en el método del valor crítico y $z_{0.05} = 1.645$, la regla de rechazo en esta prueba de la cola inferior es

$$\text{Rechazar } H_0 \text{ si } z \leq -1.645$$

Considere una muestra de 36 baterías y que por pruebas anteriores se puede considerar que se conoce σ y que su valor es $\sigma = 12$ horas. La regla de rechazo indica que se rechazará H_0 si

$$z = \frac{\bar{x} - 120}{12/\sqrt{36}} \leq -1.645$$

Despejando $\bar{x}$ de la expresión anterior se tiene que se rechazará H_0 si

$$\bar{x} \leq 120 - 1.645\left(\frac{12}{\sqrt{36}}\right) = 116.71$$

Rechazar H_0 siempre que $\bar{x} \leq 116.71$ significa que se aceptará el pedido siempre que

$$\bar{x} > 116.71$$

Con esta información ya se puede calcular la probabilidad de cometer un error tipo II. Primero, recuerde que se comete un error tipo II cuando la verdadera media del pedido es menor a 120 horas y se decida aceptar H_0: $\mu \geq 120$. Por tanto, para calcular la probabilidad de cometer un error tipo II, se debe elegir un valor de μ menor que 120 horas. Por ejemplo, suponga que la calidad del envío es mala si la vida promedio de las baterías es $\mu = 112$ horas. Si realmente es verdad que $\mu = 112$, ¿cuál es la probabilidad de aceptar H_0: $\mu \geq 120$ y cometer así un error tipo II? Observe que es la probabilidad de que la media muestral $\bar{x}$ sea mayor que 116.71 cuando $\mu = 112$.

En la figura 9.8 se muestra la distribución muestral de $\bar{x}$ si la media es $\mu = 112$. El área sombreada en la cola superior da la probabilidad de obtener una $\bar{x} > 116.7$. Usando la distribución normal estándar se ve que para t $\bar{x} = 116.7$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{116.71 - 112}{12/\sqrt{36}} = 2.36$$

La tabla de probabilidad normal estándar indica que para $z = 2.36$, el área en la cola superior es $1.0000 - 0.9909 = 0.0091$. Entonces 0.0091 es la probabilidad de cometer un error tipo II cuando $\mu = 112$. Si se usa β para denotar la probabilidad de cometer un error tipo II, se tiene que si $\mu = 112$, $\beta = 0.0091$. Se puede concluir que si la media de la población es 112 horas, la probabilidad de cometer un error tipo II es de sólo 0.0091.

Estos cálculos se repiten con otros valores de μ menores a 120. Para cada valor de μ se obtendrán diferentes probabilidades de cometer un error tipo II. Por ejemplo, en el pedido de las

FIGURA 9.8 PROBABILIDAD DE COMETER UN ERROR TIPO II CUANDO $\mu = 112$

baterías la media de la vida útil es $\mu = 115$ horas. Como H_0 será aceptada siempre que $\bar{x} > 116.71$, el valor z obtenido con $\mu = 115$ está dado por

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{116.71 - 115}{12/\sqrt{36}} = 0.86$$

En la tabla de probabilidad normal estándar, se ve que el área en la cola superior de la distribución normal estándar que corresponde a $z = 0.86$ es $1.0000 - 0.8051 = 0.1949$. Si $\mu = 115$, la probabilidad de cometer un error tipo II es $\beta = 0.1949$.

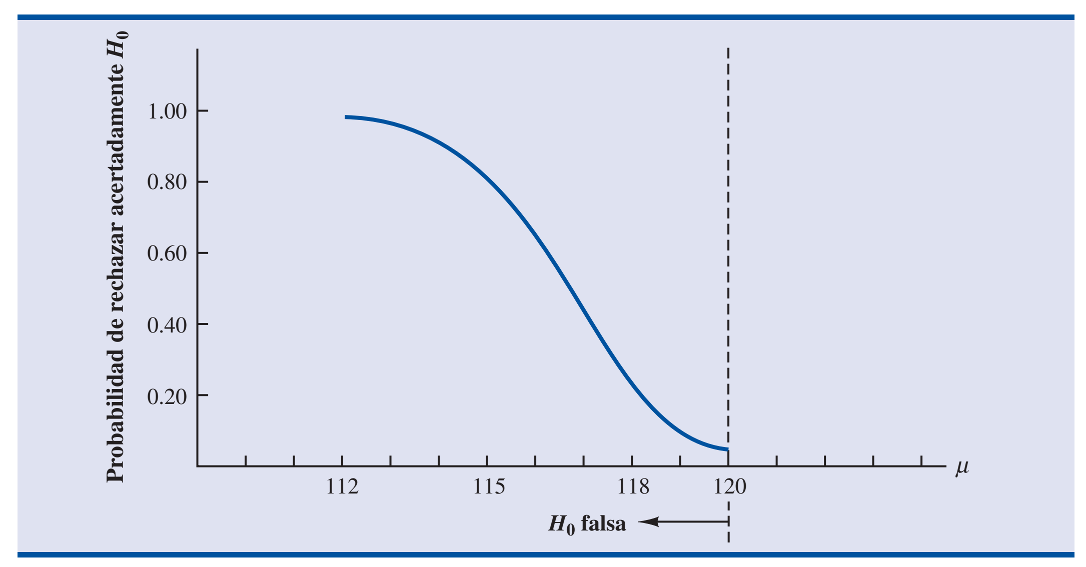
Como se muestra en la tabla 9.5, la probabilidad de cometer un error tipo II depende del valor de la media poblacional μ. Si los valores de μ son cercanos a μ_0, la probabilidad de cometer un error tipo II puede ser alta.

En la tabla 9.5 se muestran las probabilidades de cometer un error tipo II para varios valores de μ menores a 120. Observe que si μ aumenta y se acerca a 120, la probabilidad de cometer un error tipo II aumenta hacia un límite superior de 0.95. Pero a medida que μ disminuye y se aleja de 120, la probabilidad de cometer un error tipo II disminuye. Este es el patrón esperable. Cuando la verdadera media poblacional está cerca del valor de la hipótesis nula, $\mu = 120$, la probabilidad de cometer un error tipo II es grande. Pero cuando la verdadera media poblacional está lejos del valor $\mu = 120$ de la hipótesis nula, la probabilidad de cometer un error tipo II es baja.

TABLA 9.5 PROBABILIDAD DE COMETER UN ERROR TIPO II EN LA PRUEBA DE HIPÓTESIS DEL MUESTREO DE ACEPTACIÓN

Valor de μ	$z = \frac{116.71 - \mu}{12/\sqrt{36}}$	**Probabilidad de un error tipo II (β)**	**Potencia ($1 - \beta$)**
112	2.36	0.0091	0.9909
114	1.36	0.0869	0.9131
115	0.86	0.1949	0.8051
116.71	0.00	0.5000	0.5000
117	−0.15	0.5596	0.4404
118	−0.65	0.7422	0.2578
119.999	−1.645	0.9500	0.0500

FIGURA 9.9 CURVA DE POTENCIAS PARA LA PRUEBA DE HIPÓTESIS DEL MUESTREO DE ACEPTACIÓN



A la probabilidad de rechazar acertadamente H_0 cuando es falsa se le llama **potencia** de la prueba. Para cada valor de μ, la potencia es $1 - \beta$; es decir, la probabilidad de rechazar acertadamente la hipótesis nula es 1 menos la probabilidad de cometer un error tipo II. En la tabla 9.5 se presentan también los valores de la potencia. Con base en estos valores, en la figura 9.9 se presentan las potencias correspondientes a cada valor μ. A este tipo de gráficas se les conoce como **curva de potencias**. Observe que la curva de potencias se extiende sobre los valores de μ para los que la hipótesis nula es falsa. La altura en la curva de potencias para cualquier valor de μ indica la probabilidad de rechazar acertadamente H_0, cuando es falsa.*

En resumen, para calcular la probabilidad de cometer un error tipo II en una prueba de hipótesis para la media poblacional se puede seguir, paso por paso, el procedimiento siguiente.

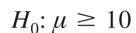
1. Formular las hipótesis nula y alternativa.
2. Usar el nivel de significancia α y el método del valor crítico para determinar el valor crítico y la regla de decisión para la prueba.
3. Usar la regla de decisión para encontrar el valor de la media muestral que corresponde al valor crítico del estadístico de prueba.
4. Usar el resultado del paso 3 para determinar el valor de la media muestral que llevará a la aceptación de H_0. Este valor define la región de aceptación de la prueba.
5. Usar la distribución muestral de $\bar{x}$ para un valor de μ que satisfaga la hipótesis alternativa y la región de aceptación del paso 4, para calcular la probabilidad de que la media muestral se encuentre en la región de aceptación. Ésta es la probabilidad de cometer un error tipo II dado el valor de μ elegido.

Ejercicios

Métodos



46. Considere la prueba de hipótesis siguiente



$$H_0: \mu \geq 10$$
$$H_a: \mu < 10$$

*Algunas veces para proporcionar información acerca de la probabilidad de cometer un error tipo II, se usa otra gráfica denominada curva característica de operación, la cual da la probabilidad de aceptar H_0, y β, con valores de μ para los que la hipótesis nula es falsa. Con esta gráfica se puede leer directamente la probabilidad de cometer un error tipo II.

El tamaño de la muestra es 120 y la desviación estándar poblacional se considera conocida, $\sigma = 5$. Use $\alpha = 0.05$.

a. Si la media poblacional es 9, ¿cuál es la probabilidad de que la media muestral lleve a la conclusión de *no rechazar* H_o?
b. ¿Qué tipo de error se comete si la media poblacional es 9 y se concluye que H_0: $\mu \geq 10$ es verdadera?
c. ¿Cuál es la probabilidad de cometer un error tipo II si la verdadera media poblacional es 8?

47. Considere la prueba de hipótesis siguiente

$$H_0: \mu = 20$$
$$H_a: \mu \neq 20$$

Se toma una muestra de 200 elementos y la desviación estándar poblacional es $\sigma = 10$. Use $\alpha = 0.05$. Calcule la probabilidad de cometer un error tipo II si la media poblacional es

a. $\mu = 18.0$
b. $\mu = 22.5$
c. $\mu = 21.0$

Aplicaciones

48. Fowle Marketing Research, Inc. tasa la cantidad que cobra a sus clientes en la suposición de que una encuesta por teléfono se puede realizar en un promedio de 15 minutos o menos. Si se necesita más tiempo en promedio, se cobra una cantidad adicional. Con una muestra de 35 encuestas, una desviación estándar de 4 minutos y 0.01 como nivel de significancia, se usará la media muestral para probar la hipótesis nula H_0: $\mu \leq 15$.

a. Dé su interpretación del error tipo II en este problema. ¿Qué impacto tiene en la empresa?
b. ¿Cuál es la probabilidad de cometer un error tipo II si la verdadera media de los tiempos es $\mu = 17$ minutos?
c. ¿Cuál es la probabilidad de cometer un error tipo II si la verdadera media de los tiempos es $\mu = 18$ minutos?
d. Dibuje la forma general de la curva de potencias de esta prueba.



49. Un grupo de investigación para los consumidores está interesado en probar la afirmación de un fabricante de automóviles de que un nuevo modelo da por lo menos 25 millas por galón de gasolina (H_0: $\mu \geq 25$).

a. Con 0.02 como nivel de significancia y una muestra de 30 automóviles, ¿cuál es la regla de rechazo basada en el valor $\bar{x}$ en la prueba para determinar si debe rechazarse la afirmación del fabricante? Suponga que σ es 3 millas por galón.
b. ¿Cuál es la probabilidad de cometer un error tipo II si el verdadero rendimiento es 23 millas por galón?
c. ¿Cuál es la probabilidad de cometer un error tipo II si el verdadero rendimiento es 24 millas por galón?
d. ¿Cuál es la probabilidad de cometer un error tipo II si el verdadero rendimiento es 25.5 millas por galón?

50. La revista *Young Adult* establece la hipótesis siguiente acerca de la edad de sus suscriptores:

$$H_0: \mu = 28$$
$$H_a: \mu \neq 28$$

a. En esta situación, ¿qué significa cometer un error tipo II?
b. Se supone que la desviación estándar muestral es $\sigma = 6$ años y que el tamaño de la muestra es 100. Si $\alpha = 0.05$, ¿cuál es la probabilidad de aceptar H_0 si μ es igual a 26, 27, 29 y 30?
c. ¿Cuál es la potencia si $\mu = 26$? ¿Qué le dice este resultado?

51. En la operación de una línea de producción se prueba que se llene con el peso exacto mediante la prueba de hipótesis siguiente.

Hipótesis	**Conclusión y medida**
H_0: $\mu = 16$	Llenado correcto; puede continuar
H_a: $\mu \neq 16$	Llenado fuera del estándar; detener y ajustar la máquina

El tamaño de la muestra es 30 y la desviación estándar poblacional es $\sigma = 0.8$. Use $\alpha = 0.05$.
a. En esta situación, ¿qué significa un error tipo II?
b. ¿Cuál es la probabilidad de cometer un error tipo II si se está llenando con 0.5 onzas de exceso?
c. Si se está llenando con 0.5 onzas de exceso, ¿cuál es la potencia de la prueba estadística?
d. Dé la curva de potencias para esta prueba estadística. ¿Qué información aporta al gerente de producción?

52. Vaya al ejercicio 48. Suponga que la empresa toma una muestra de 50 encuestas y repite los incisos b y c. ¿Qué observación cabe hacer sobre cómo afecta el tamaño de la muestra a la probabilidad de cometer un error tipo II?

53. Sparr Investments, Inc. se especializa en oportunidades de inversión para sus clientes con pago de impuestos deferido. Recién ofreció un programa de inversión con deducción de la nómina para los empleados de una determinada empresa. Sparr estima que en este momento los empleados tienen \$100 o menos por mes en inversiones con impuestos diferidos. Para probar la hipótesis de Sparr acerca de las inversiones entre la población de empleados, se toma una muestra de 40 empleados. Suponga que las cantidades invertidas mensualmente por los empleados en inversiones con impuestos diferidos tienen una desviación estándar de \$75 y que en esta prueba de hipótesis se usará 0.05 como nivel de significancia.
a. En esta situación, ¿cuál es el error tipo II?
b. ¿Cuál es la probabilidad de cometer un error tipo II si la media de la inversión mensual de los empleados es \$120?
c. ¿Cuál es la probabilidad de cometer un error tipo II si la media de la inversión mensual de los empleados es \$120?
d. Suponiendo que se usa una muestra de 80 empleados, repita los incisos b y c.



Determinación del tamaño de la muestra en una prueba de hipótesis para la media poblacional

Considere realizar una prueba de hipótesis para el valor de la media poblacional. El nivel de significancia elegido por el usuario determina la probabilidad de cometer un error tipo I en esta prueba. Al controlar el tamaño de la muestra, el usuario también controla la probabilidad de cometer un error tipo II. Enseguida se muestra cómo determinar el tamaño de la muestra en la prueba de hipótesis de la cola inferior para la media poblacional que se da a continuación.

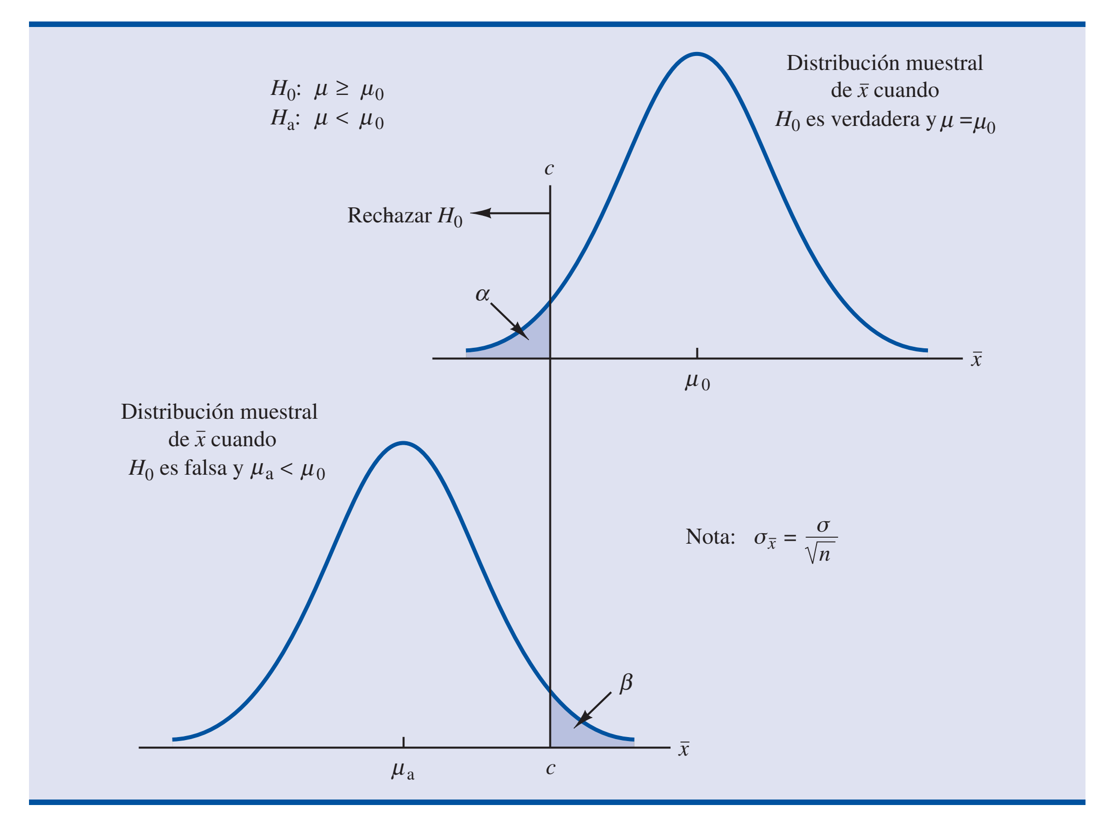
$$H_0: \mu \geq \mu_0$$
$$H_a: \mu < \mu_0$$

En la figura 9.10, la gráfica superior muestra la distribución muestral de $\bar{x}$ cuando H_0 es verdadera y $\mu = \mu_0$. En una prueba de la cola inferior el valor crítico del estadístico de prueba se denota $-z_\alpha$; la línea vertical, c, en la gráfica superior de la figura, señala el valor de $\bar{x}$ correspondiente a $-z_\alpha$. Observe que si se rechaza H_0 cuando $\bar{x} \leq c$, la probabilidad de cometer un error tipo I será α. Si z_α representa el valor de z que corresponde al área α en la cola superior de la distribución normal estándar, la fórmula siguiente se emplea para calcular c:

$$c = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} \tag{9.5}$$



FIGURA 9.10 DETERMINACIÓN DEL TAMAÑO DE LA MUESTRA PARA VALORES DADOS DE LAS PROBABILIDADES DE COMETER UN ERROR TIPO I (α) Y UN ERROR TIPO II (β)



La gráfica inferior de la figura 9.10 es la distribución muestral de $\bar{x}$ cuando la hipótesis alternativa siendo $\mu = \mu_a < \mu_0$ es verdadera. La región sombreada muestra β, la probabilidad de cometer un error tipo II al cual está expuesta la persona que toma la decisión si acepta la hipótesis nula cuando $\bar{x} > c$. Si z_β representa el valor z que corresponde al área β en la cola superior de la distribución normal estándar, c se calcula empleando la fórmula siguiente.

$$c = \mu_a + z_\beta \frac{\sigma}{\sqrt{n}} \tag{9.6}$$

Ahora lo que se busca es elegir un valor para c de manera que cuando se rechace H_0 y se acepte H_a, la probabilidad de cometer un error tipo I sea igual a la probabilidad elegida para α y la probabilidad de cometer un error tipo II sea igual al valor elegido para β. Por consiguiente, con ambas ecuaciones (9.5) y (9.6) se debe obtener el mismo valor de c y la ecuación siguiente debe satisfacerse.

$$\mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} = \mu_a + z_\beta \frac{\sigma}{\sqrt{n}}$$

Para determinar el tamaño de muestra que se necesita, primero se despeja $\sqrt{n}$ como sigue.

$$\mu_0 - \mu_a = z_\alpha \frac{\sigma}{\sqrt{n}} + z_\beta \frac{\sigma}{\sqrt{n}}$$

$$\mu_0 - \mu_a = \frac{(z_\alpha + z_\beta)\sigma}{\sqrt{n}}$$

y

$$\sqrt{n} = \frac{(z_\alpha + z_\beta)\sigma}{(\mu_0 - \mu_\text{a})}$$

Al elevar al cuadrado ambos lados de la expresión se obtiene la fórmula siguiente para el tamaño de la muestra necesario en una prueba de hipótesis de una cola para la media poblacional.

TAMAÑO DE LA MUESTRA EN UNA PRUEBA DE HIPÓTESIS DE UNA COLA PARA LA MEDIA POBLACIONAL

$$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_0 - \mu_\text{a})^2} \tag{9.7}$$

donde

z_α = valor de z que deja un área α en la cola superior de la distribución normal estándar.
z_β = valor de z que deja un área β en la cola superior de la distribución normal estándar.
σ = desviación estándar poblacional
μ_0 = valor de la media poblacional en la hipótesis nula
μ_a = valor de la media poblacional usada para el error tipo II

Nota: Para una prueba de hipótesis de dos colas, en la ecuación (9.7) se usa $z_{\alpha/2}$ en lugar de z_α.

Aunque la ecuación (9.7) se dedujo para la prueba de hipótesis de la figura 9.10, también es válida para cualquier prueba de hipótesis de una cola para la media poblacional. En una prueba de hipótesis de dos colas para la media poblacional, se usa $z_{\alpha/2}$ en lugar de z_α en la ecuación (9.7)

De regreso al ejemplo del muestreo de aceptación presentado en las secciones 9.6 y 9.7, las especificaciones para las baterías indican que la media del tiempo de vida debe ser por lo meno 120 horas. Los pedidos se regresan si se rechaza H_0: $\mu \geq 120$. Suponga que el gerente de control de calidad establece lo siguiente acerca de las probabilidades de cometer los errores tipo I y tipo II:

Para el error tipo I: si la media de la vida útil de las baterías del pedido es $\mu = 120$, estoy dispuesto a correr el riesgo de que la probabilidad de rechazar el envío sea 0.05.

Para el error tipo II: si la media de la vida útil de las baterías del pedido es 5 horas menos de lo que indican las especificaciones (es decir, $\mu = 115$), estoy dispuesto a correr el riesgo de que la probabilidad de aceptar el envío sea $\beta = 0.10$.

Lo anterior es establecido por el gerente de control de calidad con base en su propio criterio. Otra persona bien puede establecer otros valores para dichas probabilidades. Pero, tales probabilidades deben establecerse antes de determinar el tamaño de la muestra.

En el ejemplo presente se tiene, $\alpha = 0.05$ y $\beta = 0.10$. Mediante la distribución de probabilidad normal estándar se tiene $z_{0.05} = 1.645$ y $z_{0.10} = 1.28$. De acuerdo con lo dicho al especificar las probabilidades para los errores, se observa que $\mu_0 = 120$ y $\mu_\text{a} = 115$. Por último, se supuso que la desviación estándar poblacional se conocía y que era $\sigma = 12$. Mediante la ecuación (9.7), se encuentra que el tamaño de muestra recomendado para el ejemplo del muestreo de aceptación es

$$n = \frac{(1.645 + 1.28)^2(12)^2}{(120 - 115)^2} = 49.3$$

Al redondear hacia arriba, el tamaño de muestra recomendado es 50.

Como las probabilidades de los dos errores tipo I y tipo II se han controlado usando $n = 50$, queda justificado que, en esta prueba de hipótesis, el gerente de control de calidad diga *se acepta* H_0 o *se rechaza* H_0. Las inferencias correspondientes se hacen teniendo probabilidades admitidas de cometer un error tipo I o un error tipo II.

Acerca de la relación entre α, β y el tamaño n de la muestra caben tres observaciones.

1. Una vez que se tienen dos de estos tres valores, el tercero puede calcularse.
2. Dado un nivel de significancia α, aumentando el tamaño de la muestra se reduce β.
3. Dado un tamaño de muestra, al disminuir α aumenta β y al aumentar α, *disminuye* β.

La tercera observación debe tenerse en cuenta cuando no se controla la probabilidad de cometer un error tipo II. Dicha observación indica que no se deben elegir niveles de significancia α innecesariamente pequeños; para un tamaño de muestra dado, elegir un nivel de significancia pequeño implica más riesgo de cometer un error tipo II. Personas con poca experiencia piensan que al realizar una prueba de hipótesis es mejor usar valores pequeños de α. Valores pequeños de α son mejores si sólo preocupa cometer un error tipo I. Pero valores pequeños de α tienen la desventaja de incrementar la probabilidad de cometer un error tipo II.

Ejercicios

Métodos



54. Considere la prueba de hipótesis siguiente.

$$H_0: \mu \geq 10$$
$$H_a: \mu < 10$$

El tamaño de la muestra es 120 y la desviación estándar poblacional es 5. Use $\alpha = 0.05$. Si la media poblacional real es 9, la probabilidad de cometer un error tipo II es 0.2912. Suponga que el investigador desea reducir a 0.10 la probabilidad de cometer un error tipo II si la media poblacional real es 9. ¿Qué tamaño de muestra se recomienda?

55. Considere la prueba de hipótesis siguiente.

$$H_0: \mu = 20$$
$$H_a: \mu \neq 20$$

La desviación estándar poblacional es 10. Use $\alpha = 0.05$. ¿De qué tamaño deberá tomarse la muestra si el investigador está dispuesto a aceptar una probabilidad de 0.05 de cometer un error tipo II cuando la verdadera media poblacional sea 22?

Aplicaciones

56. Suponga que el director de proyecto del estudio de Hilltop Coffee (véase sección 9.3) solicita una probabilidad de 0.10 de declarar que Hilltop Coffee no cometía una violación si en realidad está llenando con 1 onza de menos ($\mu_a = 2.9375$ libras) ¿Cuál será el tamaño de muestra recomendado?



57. Una batería industrial especial debe tener una vida de por lo menos 400 horas. Considere una prueba de hipótesis con 0.02 como nivel de significancia. Si en las baterías de un determinado lote de producción la media de la vida útil es 385 horas, el gerente de producción desea un procedimiento de muestreo que sólo 10% de las veces muestre que el resultado erróneo del lote es aceptable. ¿Qué tamaño de muestra se recomienda para esta prueba de hipótesis? Use 30 horas como estimación de la desviación estándar poblacional.

58. La revista *Young Adult* plantea la hipótesis siguiente acerca de la edad de sus suscriptores.

$$H_0: \mu = 28$$
$$H_a: \mu \neq 28$$

Si el gerente que realiza la prueba admite una probabilidad de 0.15 de cometer un error tipo II si la verdadera edad promedio es 29 años. ¿De qué tamaño debe tomarse la muestra? Suponga que $\sigma = 6$ y que el nivel de significancia es 0.05.

59. En un estudio sobre el rendimiento de la gasolina se probaron las hipótesis siguientes.

Hipótesis	**Conclusión**
$H_0: \mu \geq 25$ mpg	Confirma lo que sostiene el fabricante
$H_a: \mu < 25$ mpg	Refuta lo que sostiene el fabricante, el rendimiento es menor a lo afirmado

Para $\sigma = 3$ y un nivel de significancia de 0.02, ¿qué tamaño de muestra se recomienda si el investigador desea tener 80% de posibilidad de detectar que μ es menor que 25 millas por galón, cuando realmente es 24?

Resumen

Las pruebas de hipótesis son un procedimiento estadístico que usa datos muestrales para determinar si una afirmación acerca del valor de un parámetro poblacional debe o no rechazarse. Como hipótesis se tienen dos afirmaciones opuestas acerca de un parámetro poblacional. A una de las afirmaciones se le llama hipótesis nula (H_0) y a la otra, hipótesis alternativa (H_a). En la sección 9.1 se proporcionaron los lineamientos para elaborar estas hipótesis para tres situaciones encontradas a menudo en la práctica.

Si se tienen datos históricos o alguna otra información, éstos proporcionan una base para suponer que se conoce la desviación estándar poblacional, el procedimiento de prueba de hipótesis para la media poblacional se sustenta en la distribución normal estándar. Si no se conoce σ, se usa la desviación estándar muestral s para estimar σ y el procedimiento de la prueba de hipótesis se basa en la distribución t. En ambos casos, la calidad de los resultados depende tanto de la forma de la distribución de la población como del tamaño de la muestra. Si la población tiene distribución normal, los dos procedimientos para la prueba de hipótesis son aplicables, aun con tamaños de muestra pequeños. Si la población no está distribuida normalmente, se necesitan tamaños de muestra mayores. En las secciones 9.3 y 9.4 se proporcionaron los lineamientos generales para el tamaño de la muestra. En el caso de pruebas de hipótesis para la proporción poblacional, en el procedimiento de la prueba de hipótesis se usa un estadístico de prueba sustentado en la distribución normal estándar.

En todos los casos el valor del estadístico de prueba se usa para calcular un valor-p para la prueba. Un valor-p es una probabilidad que se usa para determinar si se rechaza o no la hipótesis nula. Si el valor-p es menor o igual que el nivel de significancia α, la hipótesis nula puede rechazarse.

Las conclusiones de una prueba de hipótesis también pueden obtenerse comparando el valor del estadístico de prueba con el valor crítico. En pruebas de la cola inferior, la hipótesis nula se rechaza si el valor del estadístico de prueba es menor o igual que el valor crítico. En pruebas de la cola superior, la hipótesis nula se rechaza si el valor del estadístico de prueba es mayor o igual al valor crítico. En las pruebas de las dos colas hay dos valores críticos: uno en la cola inferior de la distribución muestral y otro en la cola superior de la distribución muestral. En este caso, la hipótesis nula se rechaza si el valor del estadístico de prueba es menor o igual al valor crítico de la cola inferior o mayor o igual que el valor crítico de la cola superior.

También se presentaron extensiones de los procedimientos de prueba de hipótesis para incluir un análisis del error tipo II. En la sección 9.7 se mostró la forma de calcular la probabilidad de cometer un error tipo II. En la sección 9.8, cómo determinar el tamaño de la muestra de manera que se controlen tanto la probabilidad de cometer un error tipo I como un error tipo II.

Glosario

Hipótesis nula Hipótesis que en una prueba de hipótesis se supone tentativamente verdadera.
Hipótesis alternativa Hipótesis que se concluye verdadera cuando se rechaza la hipótesis nula.
Error tipo I El error de rechazar H_0 cuando es verdadera.
Error tipo II El error de aceptar H_0 cuando es falsa.
Nivel de significancia Probabilidad de cometer un error tipo I cuando la hipótesis nula es verdadera como igualdad.
Prueba de una cola Prueba de hipótesis en la que debido a un valor del estadístico de prueba en una de las colas de la distribución muestral se rechaza la hipótesis nula.
Estadístico de prueba Un estadístico cuyo valor ayuda a determinar si se rechaza la hipótesis nula.
Valor-*p* Probabilidad que proporciona una medida de la evidencia, dada por la muestra, contra la hipótesis nula. Entre menor sea un valor-*p*, mayor será la evidencia contra la H_0. En una prueba de la cola inferior, el valor-*p* es la probabilidad de obtener, para el estadístico de prueba, un valor tan pequeño o menor que el proporcionado por la muestra. En una prueba de la cola superior, el valor-*p* es la probabilidad de obtener, para el estadístico de prueba, un valor tan grande o mayor que el proporcionado por la muestra. En una prueba de dos colas, el valor-*p* es la probabilidad de obtener para el estadístico de prueba un valor tan poco, o menos, probable que el proporcionado por la muestra.
Valor crítico Un valor que se compara con el estadístico de prueba para determinar si se rechaza H_0.
Prueba de dos colas Prueba de hipótesis en la que se rechaza la hipótesis nula debido a un valor del estadístico de prueba que se encuentra en cualquiera de las dos colas de la distribución muestral.
Potencia La probabilidad de rechazar adecuadamente H_0 cuando es falsa.
Curva de potencias Gráfica que da la probabilidad de rechazar H_0 para cada uno de los posibles valores del parámetro poblacional que no satisfaga la hipótesis nula. La curva de potencias proporciona las probabilidades de rechazar correctamente la hipótesis nula.

Fórmulas clave

Estadístico de prueba en una prueba de hipótesis para la media poblacional: σ conocida

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{9.1}$$

Estadístico de prueba en una prueba de hipótesis para la media poblacional: σ desconocida

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{9.2}$$

Estadístico de prueba en una prueba de hipótesis para la proporción poblacional

$$z = \frac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} \tag{9.4}$$

Tamaño de la muestra en una prueba de hipótesis de una cola para la media poblacional

$$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_0 - \mu_a)^2} \tag{9.7}$$

En una prueba de las dos colas se sustituye z_α por $z_{\alpha/2}$

Ejercicios complementarios

60. En una línea de producción el peso promedio con que se llena cada recipiente es 16 onzas. Un exceso o una falta de llenado ocasionan problemas serios y cuando se detectan es necesario que el operador detenga la línea de producción para reajustar el mecanismo de llenado. De acuerdo con datos del pasado se supone que la desviación estándar poblacional es $\sigma = 0.8$ onzas. Cada hora, un inspector de producción toma muestras de 30 recipientes y decide si es necesario detener la producción y hacer un reajuste. El nivel de significancia es $\alpha = 0.05$.
    a. Establezca la prueba de hipótesis para esta aplicación al control de calidad.
    b. Si se encuentra que la media muestral es $\bar{x} = 16.32$ onzas, ¿cuál es el valor-*p*? ¿Qué medidas recomendaría usted tomar?
    c. Si se encuentra que la media muestral es $\bar{x} = 15.82$ onzas, ¿cuál es el valor-*p*? ¿Qué medidas recomendaría usted tomar?
    d. Use el método del valor crítico. ¿Cuál es la regla de decisión en la prueba de hipótesis anterior? Repita los incisos b y c ¿Llega a la misma conclusión?

61. En la Western University, la media histórica poblacional en las puntuaciones de los solicitantes de una beca es 900. La desviación estándar poblacional histórica que se considera conocida es $\sigma = 180$. Cada año se toma una muestra de los solicitantes para determinar si esta media ha cambiado.
    a. Establezca las hipótesis
    b. ¿Cuál es el intervalo de 95% de confianza para la estimación de la media poblacional de las puntuaciones en el examen, si en una muestra de 200 estudiantes la media muestral es $\bar{x} = 935$?
    c. Use el intervalo de confianza para realizar una prueba de hipótesis. Usando $\alpha = 0.05$, ¿a qué conclusión llega?
    d. ¿Cuál es el valor-*p*?

62. *Playbill* es una revista que se distribuye entre las personas que asisten a conciertos y al teatro. El ingreso medio anual por familia en la población de lectores de *Playbill* es $119 155 (*Playbill*, enero de 2006). Suponga que la desviación estándar es $\sigma = \$20\,700$. Un grupo de San Francisco asegura que entre las personas de la zona de la Bahía que van al teatro el ingreso medio es más alto. En una muestra de 60 personas de la Bahía que suelen ir al teatro, el ingreso medio por hogar fue $126 100.
    a. Establezca las hipótesis para determinar si los datos muestrales apoyan la conclusión de que las personas de la zona de la Bahía que suelen asistir al teatro tienen un ingreso medio por familia más alto que los demás lectores de *Playbill*.
    b. ¿Cuál es el valor-*p* a partir de la muestra de las 60 personas de la Bahía que suelen ir al teatro?
    c. Use $\alpha = 0.01$ como nivel de significancia. ¿A qué conclusión llega?

63. El viernes los corredores de bolsa de Wall Street esperaban ansiosos la publicación del gobierno federal del aumento en enero de nóminas no agrícolas. El primer consenso estimado entre los economistas fue que se esperaba un aumento de 250 000 nuevos empleos (CNBC, 3 de febrero de 2006). Sin embargo, en una muestra de 20 economistas tomada el jueves en la tarde, la media muestral fue 266 000, y la desviación estándar muestral 24 000. Los analistas financieros suelen llamar a tales medias muestrales, basadas en las estimaciones que circulan en el mercado después de que los analistas incorporan las últimas informaciones, "whisper number". Realice una prueba de hipótesis para determinar si el "whisper number" justifica la conclusión de un aumento estadísticamente significante en la estimación de consenso de los economistas. Use $\alpha = 0.01$ como nivel de significancia.

64. El consejo universitario informa que el número promedio de estudiantes de nuevo ingreso en las universidades es 6000 (*USA Today*, 26 de diciembre de 2002). En un periodo reciente de inscripciones se tomó una muestra de 32 universidades con una media muestral de los estudiantes de nuevo ingreso de 5812 y una desviación estándar muestral de 1140. ¿Estos datos indican un cambio en el número medio de estudiantes de nuevo ingreso? Use $\alpha = 0.05$.

65. En un estudio sobre los costos de atención a la salud en Estados Unidos se presentaron datos que mostraban un gasto medio de Medicare por derechohabiente de $6 883 en el 2003. Para investigar las diferencias en todo el país, un investigador tomó una muestra de 40 derechohabientes de

Medicare del estado de Indianápolis. En la muestra de Indianápolis el gasto promedio de Medicare en el 2003 fue $5980 y la desviación estándar $2518.

a. Establezca las hipótesis a usar si se quiere determinar si el gasto anual medio de Medicare en Indianápolis es menor a la media nacional.
b. Use los resultados muestrales anteriores para calcular el estadístico de prueba y el valor-p.
c. Use $\alpha = 0.05$. ¿A qué conclusión llega?
d. Repita la prueba de hipótesis usando el método del valor crítico.

66. La cámara de comercio de una comunidad de Florida anuncia en su publicidad que el costo medio de un terreno residencial es $125 000 o menos por lote. Suponga que en una muestra de 32 lotes se encuentra que la media muestral es $130 000 por lote y que la desviación estándar muestral es $12 500. Use 0.05 como nivel de significancia para probar la validez de lo que se dice en la publicidad.



67. La U.S. Energy Administration informó que en Estados Unidos el precio medio del galón de gasolina era $2.357 (U.S. Energy Administration, 30 de enero de 2006). En el archivo llamado Gasoline se encuentran los precios de gasolina normal encontrados en una muestra de 50 gasolineras en estados del Atlántico sur. Realice una prueba de hipótesis para determinar si el precio medio del galón de gasolina en los estados del Atlántico sur es diferente a la media nacional. Use $\alpha = 0.05$ como nivel de significancia y dé su conclusión.

68. En un estudio del Center of Disease Control, CDC, se encontró que 23% de los adultos son fumadores y de éstos 70% indicaron que quieren dejar de fumar (Associated Press, 22 de julio de 2002). El CDC informó que, quienes fumaron en algún momento de su vida, 50% habían podido dejar ese hábito. Parte del estudio indicó que el éxito en dejar de fumar aumenta con el nivel de estudios. Suponga que en una muestra de 100 personas con título universitario que han fumado en algún momento de su vida, 64 lograron dejar de fumar.

a. Establezca las hipótesis a usar para determinar si la población de personas con título universitario tiene más éxito para dejar de fumar que la población general.
b. Dados los datos muestrales, ¿cuál es la proporción de personas con título universitario, que habiendo fumado en algún momento de su vida, pudieran dejar de hacerlo?
c. ¿Cuál es el valor-p? Con $\alpha = 0.01$, ¿cuál es la conclusión de la prueba de hipótesis?

69. La promoción de una línea aérea se sustenta en la suposición de que dos terceras partes de los viajeros de negocios usan una computadora portátil en sus viajes de negocios durante la noche.

a. Establezca las hipótesis a usar para probar esta suposición.
b. ¿Cuál es la proporción muestral encontrada en un estudio patrocinado por American Express, en que 355 de 546 viajeros de negocios utilizaban su computadora portátil en sus viajes de negocios durante la noche?
c. ¿Cuál es el valor-p?
d. Use $\alpha = 0.05$. ¿A qué conclusión llega?

70. Los centros virtuales de llamadas son atendidos por personas que trabajan fuera de sus hogares. La mayor parte de los agentes de casa (home agent) ganan $10 a $15 por hora sin beneficios frente a $7 a $9 por hora con beneficios en un centro tradicional de llamadas (*BusinessWeek*, 23 de enero de 2006). Regional Airways está considerando emplear agentes de casa, pero sólo si conservan una satisfacción del cliente mayor que 80%. Se realizó una prueba con agentes de casa. En una muestra de 300 clientes, 252 indicaron estar satisfechos con el servicio.

a. Elabore las hipótesis de una prueba para determinar si los datos muestrales apoyan la conclusión de que el servicio al cliente con agentes de casa satisface el criterio de Regional Airways.
b. ¿Cuál es la estimación puntual del porcentaje de clientes satisfechos?
c. ¿Cuál es el valor-p proporcionado por los datos muestrales?
d. ¿Cuál es la conclusión en esta prueba de hipótesis? Use como nivel de significancia $\alpha = 0.05$.

71. Durante el año de elecciones 2004, diario se publicaban los resultados de los nuevos sondeos. En un sondeo de IBD/TIPP de 910 entrevistados, 503 dijeron sentirse optimistas ante las perspectivas nacionales y el índice de liderazgo del presidente Bush subió de 4.7 puntos a 55.3 (*Investor´s Business Daily*, 14 de enero de 2004).

a. ¿Cuál es la proporción muestral de encuestados optimistas ante las perspectivas nacionales?
b. Un director de campaña quiere afirmar que el sondeo indica que la mayoría de los adultos se sienten optimistas ante las perspectivas nacionales. Elabore una prueba de hipótesis de manera que el rechazo de la hipótesis nula permita concluir que la proporción de optimistas es mayor a 50%.
c. Use los datos del sondeo para calcular el valor-*p* en la prueba de hipótesis del inciso b. Explique al director lo que dice este valor-*p* acerca del nivel de significancia de los resultados.

72. La estación de radio de Myrtle Beach, una localidad vacacional, anuncia que 90% de los hoteles estarán llenos el fin de semana en que se conmemora el Memorial Day. Dicha estación de radio aconseja a sus oyentes hacer sus reservaciones con anticipación si piensan pasar ese fin de semana en esa localidad. La noche del sábado, en una muestra de 58 hoteles, 49 estaban completamente llenos y 9 aún tenían habitaciones libres. ¿Cuál es su reacción a lo anunciado por la estación de radio después de ver las evidencias muestrales? Use $\alpha = 0.05$ en esta prueba estadística. ¿Cuál es el valor-*p*?

73. En Estados Unidos, de acuerdo con el gobierno federal, 24% de los trabajadores amparados por el plan de atención a la salud de su empresa no tuvieron que contribuir a la prima (*Statistical Abstract of the United States: 2006*). En un estudio reciente se encontró que a 81 trabajadores de los 400 tomados en una muestra no se les pidió que contribuyeran para el plan de atención a la salud de su empresa.
a. Elabore las hipótesis para probar si ha disminuido el porcentaje de trabajadores a quienes no se les pide que contribuyan para el plan de atención a la salud de su empresa.
b. ¿Cuál es la estimación puntual de la proporción que tiene un seguro de salud financiado totalmente por su empresa?
c. ¿Ha habido una disminución estadísticamente significativa en la proporción de trabajadores que tienen un seguro de salud financiado totalmente por su empresa? Use $\alpha = 0.05$.

74. Shorney Construction Company licita proyectos suponiendo que el tiempo desperdiciado por trabajador es menos de 72 minutos por día. Para probar esta suposición se usa una muestra de 30 trabajadores de la construcción. Suponga que la desviación estándar poblacional es 20 minutos.
a. Establezca las hipótesis para esta prueba.
b. ¿Cuál es la probabilidad de cometer un error tipo II si la media poblacional fueran 80 minutos?
c. ¿Cuál es la probabilidad de cometer un error tipo II si la media poblacional fueran 75 minutos?
d. ¿Cuál es la probabilidad de cometer un error tipo II si la media poblacional fueran 70 minutos?
e. Bosqueje la curva de potencias para este problema.

75. Existe un programa de ayuda federal para las zonas de bajos ingresos. Para recibir esta ayuda, el ingreso medio de la zona debe ser menor que $15 000 anuales. Zonas en las que el ingreso medio anual sea $15 000 o más no pueden recibir esta ayuda. Para decidir si una zona recibe la ayuda, se toma una muestra de los habitantes de esa zona y se realiza una prueba de hipótesis con 0.02 como nivel de significancia. Si los lineamientos establecen una probabilidad máxima de 0.05 de no dar esta ayuda a una zona en la que el ingreso medio anual sea de $14 000, ¿qué tamaño de muestra deberá usarse en el estudio? Use $\sigma = \$4000$.

76. Para probar si en la producción de un jabón de baño se satisface el estándar de producir 120 barras por lote se usan las hipótesis H_0: $\mu = 120$ y H_a: $\mu \neq 120$. Use 0.05 como nivel de significancia en esta prueba y 5 para la desviación estándar.
a. Si la media de producción llega a 117 barras por lote, la empresa desea tener 98% de oportunidad de concluir que no se está satisfaciendo el estándar de producción. ¿De qué tamaño deberá tomarse la muestra?
b. Con el tamaño de muestra del inciso a, ¿cuál es la probabilidad de concluir que el proceso está operando insatisfactoriamente si la media de producción real es: 117, 118, 119, 121, 122 y 123 barras por lote? Es decir, ¿cuál es, en cada caso, la probabilidad de cometer un error tipo II?

Caso problema 1 Quality Associates, Inc.

Quality Associates, Inc., una empresa consultora, asesora a sus clientes acerca de procedimientos estadísticos y de muestreo para el control de sus procesos de fabricación. En una determinada asesoría, el cliente dio a Quality Associates una muestra de 800 observaciones tomadas mientras el proceso del cliente operaba satisfactoriamente. La desviación estándar de estos datos fue 0.21; al ser tantos los datos, se consideró que la desviación estándar poblacional era 0.21. Quality Associates recomendó que para vigilar el proceso periódicamente se tomaran muestras aleatorias de tamaño 30. Al analizar estas muestras, el cliente sabrá pronto si el proceso era adecuado. Si el proceso no lo era, se podían tomar las medidas necesarias para eliminar el problema. De acuerdo con las especificaciones, la media en el proceso debería ser 12. A continuación, la prueba de hipótesis sugerida por Quality Associates.

$$H_0: \mu = 12$$
$$H_a: \mu \neq 12$$

Siempre que se rechazara H_0 deberían tomarse las medidas adecuadas.

Durante el primer día en que se realizó este nuevo proceso estadístico de control se tomaron las siguientes muestras a intervalos de una hora. Estos datos se encuentran en el conjunto de datos Quality

archivo en CD
Quality

Muestra 1	Muestra 2	Muestra 3	Muestra 4
11.55	11.62	11.91	12.02
11.62	11.69	11.36	12.02
11.52	11.59	11.75	12.05
11.75	11.82	11.95	12.18
11.90	11.97	12.14	12.11
11.64	11.71	11.72	12.07
11.80	11.87	11.61	12.05
12.03	12.10	11.85	11.64
11.94	12.01	12.16	12.39
11.92	11.99	11.91	11.65
12.13	12.20	12.12	12.11
12.09	12.16	11.61	11.90
11.93	12.00	12.21	12.22
12.21	12.28	11.56	11.88
12.32	12.39	11.95	12.03
11.93	12.00	12.01	12.35
11.85	11.92	12.06	12.09
11.76	11.83	11.76	11.77
12.16	12.23	11.82	12.20
11.77	11.84	12.12	11.79
12.00	12.07	11.60	12.30
12.04	12.11	11.95	12.27
11.98	12.05	11.96	12.29
12.30	12.37	12.22	12.47
12.18	12.25	11.75	12.03
11.97	12.04	11.96	12.17
12.17	12.24	11.95	11.94
11.85	11.92	11.89	11.97
12.30	12.37	11.88	12.23
12.15	12.22	11.93	12.25

Informe administrativo

1. Con cada una de las muestras realice una prueba de hipótesis usando 0.01 como nivel de significancia, determine las medidas a tomar. Dé el estadístico de prueba y el valor-p de cada prueba.
2. Calcule la desviación estándar de cada una de las cuatro muestras. ¿Parece razonable considerar 0.21 como la desviación estándar poblacional?
3. Calcule límites alrededor de $\mu = 12$ para la media muestral $\bar{x}$ de manera que, en tanto las medias muestrales se encuentren dentro de estos límites, pueda considerarse que el proceso opera adecuadamente. Pero si $\bar{x}$ es mayor al límite superior o menor al límite inferior, tomar las medidas correctivas será necesario. Estos límites se conocen en el control de calidad como límites de control superior e inferior.
4. Analice las consecuencias de cambiar el nivel de significancia por un valor mayor. ¿Qué error crece si se aumenta el valor del nivel de significancia?

Caso problema 2 Estudio sobre el desempleo

La U.S. Bureau of Labor Statistics publica cada mes diversas estadísticas sobre el desempleo, entre éstas se encuentran el número de personas desempleadas y la duración media del desempleo. En noviembre de 1998, dicha oficina informó que la media nacional de tiempo de desempleo eran 14.6 semanas.

El alcalde de Filadelfia solicitó un estudio sobre la situación del desempleo en su alcaldía. En una muestra de 50 habitantes de Filadelfia se incluyó su edad y el número de semanas que estuvieron desempleados. A continuación se presenta una parte de los datos recogidos. El conjunto de datos completo se encuentra en el archivo BLS



BLS

Edad	Semanas	Edad	Semanas	Edad	Semanas
56	22	22	11	25	12
35	19	48	6	25	1
22	7	48	22	59	33
57	37	25	5	49	26
40	18	40	20	33	13

Informe administrativo

1. Use la estadística descriptiva para resumir estos datos.
2. Elabore un intervalo de confianza de 95% para estimar la media de la edad de los individuos desempleados en Filadelfia.
3. Realice una prueba de hipótesis para determinar si la media de la duración del desempleo en Filadelfia es mayor que la media nacional, que es 14.6 semanas. Como nivel de significancia use 0.01. ¿Cuál es la conclusión?
4. ¿Hay alguna relación entre la edad y la duración del desempleo de un individuo? Explique.

Apéndice 9.1 Pruebas de hipótesis con Minitab

Se describe el uso de Minitab para realizar pruebas de hipótesis para la media poblacional y para la proporción poblacional.

Media poblacional: σ conocida

Se ilustra con el ejemplo presentado en la sección 9.3 acerca de la distancia recorrida por las pelotas de golf de Max Flight. Los datos están en la columna C1 de la hoja de cálculo de Minitab. La desviación estándar poblacional conocida es $\sigma = 12$, el nivel de significancia es $\alpha = 0.05$.

Para probar la hipótesis H_0: $\mu = 295$ frente a la hipótesis H_a: $\mu \neq 295$ se siguen los pasos que se indican:



GolfTest

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Basic Statistics**
Paso 3. Elegir **1-Sample Z**
Paso 4. Cuando aparezca el cuadro de diálogo 1-Sample Z:
Ingresar C1 en el cuadro **Samples in columns**
Ingresar 12 en el cuadro **Standar deviation**
Ingresar 295 en el cuadro **Test mean**
Seleccionar **Options**
Paso 5. Cuando aparezca el cuadro de diálogo 1-Sample Z Options:
Ingresar 95 en el cuadro **Confidence level***
Seleccionar **not equal** en el cuadro **Alternative**
Clic en **OK**
Paso 6. Clic en **OK**

Además de los resultados de la prueba de hipótesis, Minitab proporciona un intervalo de confianza de 95% para la media poblacional.

Este procedimiento se modifica fácilmente para una prueba de hipótesis de una cola seleccionando la opción menor que (less than) o mayor que (greater than) en el cuadro **Alternative** del paso 5.

Media poblacional: σ desconocida



AirRating

Las puntuaciones dadas por 60 viajeros de negocios al aeropuerto de Heathrow se han ingresado en la columna C1 de la hoja de cálculo de Minitab. El nivel de significancia para esta prueba es $\alpha = 0.05$ y la desviación estándar poblacional σ se estimará mediante la desviación estándar muestral s. Para probar la hipótesis H_0: $\mu \leq 7$ frente a la hipótesis H_a: $\mu > 7$ se usan los pasos siguientes.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Basic Statistics**
Paso 3. Elegir **1-Sample t**
Paso 4. Cuando aparezca el cuadro de diálogo 1-Sample t
Ingresar C1 en el cuadro **Samples in columns**
Ingresar 7 en el cuadro **Test mean**
Seleccionar **Options**
Paso 5. Cuando aparezca el cuadro de diálogo 1-Sample t-options:
Ingresar 95 en el cuadro **Confidence level**
Seleccionar **greater than** en el cuadro **Alternative**
Clic en **OK**
Paso 6. Clic en **OK**

En el estudio de las puntuaciones para el aeropuerto de Heathrow se tiene una hipótesis alternativa mayor que. Los pasos anteriores se modifican con facilidad para otras pruebas de hipótesis seleccionando las opciones menor que (less than) o no igual (not equal) en el cuadro **Alternative** del paso 5.

Proporción poblacional



WomenGolf

Se ilustra usando el ejemplo del campo de golf Pine Creek presentado en la sección 9.5. Los datos con las respuestas mujer (Female) y hombre (Male) están en la columna C1 de la hoja de cálculo de Minitab. Minitab usa un orden alfabético de las respuestas y selecciona la *segunda respuesta* para la proporción poblacional de interés. En este caso Minitab usa el orden alfabéti-

*Minitab proporciona simultáneamente los resultados de la prueba de hipótesis y la estimación por intervalo. El usuario selecciona cualquier nivel de confianza para la estimación por intervalo de la media poblacional: aquí se sugiere 95% de confianza.

co Female-Male (mujer-hombre) y da la proporción poblacional de las respuestas Male (hombre). Como Female (mujer) es la respuesta de interés, hay que modificar el orden alfabético de Minitab como sigue. Seleccionar cualquier celda de la columna y usar la secuencia Editor > Column > Value Order. Después elegir la opción de ingresar un orden especificado por el usuario. Ingresar Male-Female en el cuadro **Define-an-order** y hacer clic en OK. Minitab 1 Proportion es la rutina que suministra los resultados de la prueba de hipótesis para la proporción poblacional de golfistas. Proceda como sigue:

Paso 1. Paso 2. Seleccionar el menú Stat
Paso 2. Elegir **Basic Statistics**
Paso 3. Paso 3. Elegir **1 Proportion**
Paso 3. Cuando aparezca el cuadro de diálogo 1 Proportion:
 Ingresar C1 en el cuadro **Samples in Columns**
 Seleccionar **Options**
Paso 5. Cuando aparezca el cuadro de diálogo 1 Proportion:
 Ingresar 95 en el cuadro **Confidence level**
 Ingresar 0.20 en el cuadro **Test proportion**
 Seleccionar **greater than** en el cuadro **Alternative**
 Seleccionar **Use test and interval based on normal distribution**
 Clic en **OK**
Paso 6. Clic en **OK**

Apéndice 9.2 Prueba de hipótesis con Excel

Excel no cuenta con rutinas predefinidas para las pruebas de hipótesis presentadas en este capítulo. Dichas limitaciones se resuelven presentando hojas de cálculo de Excel, diseñadas por los autores de este libro, para pruebas de hipótesis, la media poblacional y la proporción poblacional. Usar estas hojas de cálculo es sencillo y también pueden modificarse para cualesquiera datos muestrales. Las hojas de cálculo se encuentran también en el disco compacto que viene con el libro.

Media poblacional: σ conocida

Se ilustra con el ejemplo presentado en la sección 9.3 de la distancia de las pelotas de golf de Max Flight. Los datos están en la columna A de la hoja de cálculo de Excel. Se conoce la desviación estándar poblacional y es $\sigma = 12$, con nivel de significancia es $\alpha = 0.05$. Para probar la hipótesis H_0: $\mu = 295$ frente a la hipótesis H_a: $\mu \neq 295$ se siguen los pasos que se indican a continuación.



Hyp Sigma Known

A medida que se describe este procedimiento, consulte la figura 9.11. En la hoja de cálculo que aparece en segundo plano se muestran las celdas con las fórmulas usadas para calcular los resultados que se muestran en la hoja de cálculo en primer plano. Los datos se han introducido en las celdas A2:A51. Para usar la plantilla son necesarios los pasos siguientes.

Paso 1. Ingresar A2:A51 en la fórmula =CONTAR de la celda D4
Paso 2. Ingresar A2:A51 en la fórmula =PROMEDIO de la celda D5
Paso 3. Ingresar la desviación estándar poblacional $\sigma = 12$ en la celda D6
Paso 4. Ingresar el valor hipotético de la media poblacional 295 en la celda D8

Las fórmulas de las celdas restantes proporcionarán en automático el error estándar, el valor del estadístico de prueba z y tres valores-p. Como la hipótesis alternativa ($\mu_0 \neq 295$) indica que se trata de una prueba de dos colas, para tomar la decisión de rechazar o no, se usa el valor-p (Two Tail) de la celda D15. Como el valor-p = 0.1255 > α = 0.05, no se puede rechazar la hipótesis nula. Los valores-p de las celdas D13 y D14 se usarían si se tratara de una prueba de hipótesis de una sola cola.

FIGURA 9.11 HOJA DE CÁLCULO DE EXCEL PARA PRUEBAS DE HIPÓTESIS PARA LA MEDIA POBLACIONAL CON σ CONOCIDA

	A	B	C	D	E
1	Yards		Hypothesis Test About a Population Mean		
2	303		With σ Known		
3	282				
4	289		Sample Size	=COUNT(A2:A51)	
5	298		Sample Mean	=AVERAGE(A2:A51)	
6	283		Population Std. Deviation	12	
7	317				
8	297		Hypothesized Value	295	
9	308				
10	317		Standard Error	=D6/SQRT(D4)	
11	293		Test Statistic z	=(D5-D8)/D10	
12	284				
13	290		p-value (Lower Tail)	=NORMSDIST(D11)	
14	304		p-value (Upper Tail)	=1-D13	
15	290		p-value (Two Tail)	=2*MIN(D13,D14)	
16	311				
17	305				
49	303				
50	301				
51	292				
52					

	A	B	C	D	E
1	Yards		Hypothesis Test About a Population Mean		
2	303		With σ Known		
3	282				
4	289		Sample Size	50	
5	298		Sample Mean	297.6	
6	283		Population Std. Deviation	12	
7	317				
8	297		Hypothesized Value	295	
9	308				
10	317		Standard Error	1.70	
11	293		Test Statistic z	1.53	
12	284				
13	290		p-value (Lower Tail)	0.9372	
14	304		p-value (Upper Tail)	0.0628	
15	290		p-value (Two Tail)	0.1255	
16	311				
17	305				
49	303				
50	301				
51	292				
52					

Nota: Los renglones 18 a 48 se encuentran ocultos.

Esta plantilla se usa para los cálculos de pruebas de hipótesis de otras aplicaciones. Por ejemplo, para realizar una prueba de hipótesis con otro conjunto de datos, ingrese los datos en la columna A de la hoja de cálculo. Las fórmulas de las celdas D4 y D5 se modifican para que correspondan a las celdas en que se encuentran los datos. Para obtener los resultados ingrese la desviación estándar poblacional en la celda D6, y en la celda D8 se ingresa la media poblacional hipotética. Si los datos muestrales ya han sido resumidos, no es necesario ingresarlos en la hoja de cálculo. En este caso, para obtener los resultados se ingresa el tamaño de la muestra en la cel-

da D4, la media muestral en la celda D5, la desviación estándar poblacional en la celda D6 y el valor hipotético de la media poblacional en la celda D8. La hoja de cálculo que se presenta en la figura 9.11 se encuentra bajo el nombre Hyp Sigma Known en el disco compacto que viene con el libro.

Media poblacional: σ desconocida

Se ilustra usando el ejemplo presentado en la sección 9.4 de las puntuaciones dadas al aeropuerto de Heathrow. Los datos están en la columna A de la hoja de cálculo de Excel. La desviación estándar poblacional σ no se conoce y será estimada mediante la desviación estándar muestral s. El nivel de significancia es $\alpha = 0.05$. Para probar la hipótesis H_0: $\mu \leq 7$ frente a la hipótesis H_a: $\mu > 7$ se usan los pasos siguientes.



Consulte la figura 9.12 a medida que se describe este procedimiento. La hoja de cálculo que aparece en segundo plano muestra las fórmulas usadas para obtener los resultados en la versión en primer plano de la hoja de cálculo. Los datos se ingresan en las celdas A2:A61. Para usar la plantilla con estos datos son necesarios los pasos siguientes.



Paso 1. Ingresar A2:A61 en la fórmula =CONTAR de la celda D4
Paso 2. Ingresar A2:A61 en la fórmula =PROMEDIO de la celda D5
Paso 3. Ingresar A2:A61 en la fórmula =DESVEST de la celda D6
Paso 4. Ingresar el valor hipotético 7 de la media poblacional en la celda D8

Las fórmulas de las celdas restantes proporcionarán automáticamente el error estándar, el valor del estadístico de prueba t, el número de grados de libertad y tres valores-p. Como la hipótesis alternativa ($\mu > 7$) indica que se trata de una prueba de la cola superior, para tomar la decisión de rechazar o no, se usa el valor-p (Upper Tail) de la celda D15. Como el valor-$p = 0.0353 < \alpha = 0.05$, se rechaza la hipótesis nula. Los valores-p de las celdas D14 y D16 se usarían si se tratara de una prueba de hipótesis de la cola inferior o de una prueba de hipótesis de dos colas.

Esta plantilla se usa para los cálculos de pruebas de hipótesis de otras aplicaciones. Por ejemplo, para realizar una prueba de hipótesis con otro conjunto de datos, se ingresan los datos en la columna A de la hoja de cálculo y se modifican Las fórmulas de las celdas D4, D5 y D6 para que correspondan a las celdas en que se encuentran los datos. Para obtener los resultados se ingresa en la celda D8 el valor hipotético de la media poblacional. Si los datos muestrales ya han sido resumidos, no es necesario ingresarlos en la hoja de cálculo. En este caso, para obtener los resultados se ingresa el tamaño de la muestra en la celda D4, la media muestral en la celda D5, la desviación estándar muestral en la celda D6 y el valor hipotético de la media poblacional en la celda D8. La hoja de cálculo que se presenta en la figura 9.12 se encuentra en el archivo con el nombre Hyp Sigma Unknown en el disco compacto que viene con el libro.

Proporción poblacional



Los datos con las respuestas mujer (Female) y hombre (Male) están en la columna A de la hoja de cálculo de Excel. Consulte la figura 9.13 a medida que se describe este procedimiento. La hoja de cálculo que aparece en segundo plano muestra las fórmulas usadas para obtener los resultados que aparecen en la hoja de cálculo que está en primer plano. Los datos están en las celdas A2:A401. Para probar las hipótesis H_0: $p \leq 0.20$ frente a H_a: $p > 0.20$ se usan los pasos siguientes.

Paso 1. Ingresar A2:A401 en la fórmula =CONTARA de la celda D3
Paso 2. Ingresar Female como respuesta de interés en la celda D4
Paso 3. Ingresar A2:A401 en la fórmula =CONTAR.SI de la celda D5
Paso 4. Ingresar el valor hipotético 0.20 de la proporción poblacional en la celda D8

Las fórmulas de las celdas restantes proporcionarán automáticamente el error estándar, el valor del estadístico de prueba z y tres valores-p. Como la hipótesis alternativa ($p_0 > 0.20$) indica que

FIGURA 9.12 HOJA DE CÁLCULO DE EXCEL PARA PRUEBAS DE HIPÓTESIS PARA LA MEDIA POBLACIONAL CON σ DESCONOCIDA.

	A	B	C	D	E
1	**Rating**		**Hypothesis Test About a Population Mean**		
2	5		**With σ Unknown**		
3	7				
4	8		**Sample Size**	=COUNT(A2:A61)	
5	7		**Sample Mean**	=AVERAGE(A2:A61)	
6	8		**Sample Std. Deviation**	=STDEV(A2:A61)	
7	8				
8	8		**Hypothesized Value**	7	
9	7				
10	8		**Standard Error**	=D6/SQRT(D4)	
11	10		**Test Statistic *t***	=(D5-D8)/D10	
12	6		**Degrees of Freedom**	=D4-1	
13	7				
14	8		***p*-value (Lower Tail)**	=IF(D11<0,TDIST(-D11,D12,1),1-TDIST(D11,D12,1))	
15	8		***p*-value (Upper Tail)**	=1-D14	
16	9		***p*-value (Two Tail)**	=2*MIN(D14,D15)	
17	7				
59	7				
60	7				
61	8				
62					

	A	B	C	D	E
1	**Rating**		**Hypothesis Test About a Population Mean**		
2	5		**With σ Unknown**		
3	7				
4	8		**Sample Size**	60	
5	7		**Sample Mean**	7.25	
6	8		**Sample Std. Deviation**	1.05	
7	8				
8	8		**Hypothesized Value**	7	
9	7				
10	8		**Standard Error**	0.136	
11	10		**Test Statistic *t***	1.841	
12	6		**Degrees of Freedom**	59	
13	7				
14	8		***p*-value (Lower Tail)**	0.9647	
15	8		***p*-value (Upper Tail)**	0.0353	
16	9		***p*-value (Two Tail)**	0.0706	
17	7				
59	7				
60	7				
61	8				
62					

Nota: Los renglones 18 a 48 se encuentran ocultos.

se trata de una prueba de la cola superior, para tomar la decisión de rechazar o no, se usa el valor-p (Upper Tail) de la celda D14. Como el valor-$p = 0.0062 < \alpha = 0.05$, se rechaza la hipótesis nula. Los valores-p de las celdas D13 y D15 se usarían si se tratara de una prueba de hipótesis de la cola inferior o de dos colas.

Esta plantilla se puede usar para los cálculos de pruebas de hipótesis con otras aplicaciones. Por ejemplo, para realizar una prueba de hipótesis con otro conjunto de datos, se ingresan los datos en la columna A de la hoja de cálculo. Se modifican las fórmulas de las celdas D3 y D5 para que correspondan a las celdas en que se encuentran los datos. Para obtener los resultados se ingresa en la celda D4 la respuesta de interés y en la celda D8 el valor hipotético de la proporción

FIGURA 9.13 HOJA DE CÁLCULO DE EXCEL PARA PRUEBAS DE HIPÓTESIS PARA LA PROPORCIÓN POBLACIONAL

	A	B	C	D	E
1	**Golfer**		**Hypothesis Test About a Population Proportion**		
2	Female				
3	Male		**Sample Size**	=COUNTA(A2:A401)	
4	Female		**Response of Interest**	Female	
5	Male		**Count for Response**	=COUNTIF(A2:A401,D4)	
6	Male		**Sample Proportion**	=D5/D3	
7	Female				
8	Male		**Hypothesized Value**	0.20	
9	Male				
10	Female		**Standard Error**	=SQRT(D8*(1-D8)/D3)	
11	Male		**Test Statistic *z***	=(D6-D8)/D10	
12	Male				
13	Male		***p*-value (Lower Tail)**	=NORMSDIST(D11)	
14	Male		***p*-value (Upper Tail)**	=1-D13	
15	Male		***p*-value (Two Tail)**	=2*MIN(D13,D14)	
16	Female				
400	Male				
401	Male				
402					

	A	B	C	D	E
1	**Golfer**		**Hypothesis Test About a Population Proportion**		
2	Female				
3	Male		**Sample Size**	400	
4	Female		**Response of Interest**	Female	
5	Male		**Count for Response**	100	
6	Male		**Sample Proportion**	0.2500	
7	Female				
8	Male		**Hypothesized Value**	0.20	
9	Male				
10	Female		**Standard Error**	0.0200	
11	Male		**Test Statistic *z***	2.50	
12	Male				
13	Male		***p*-value (Lower Tail)**	0.9938	
14	Male		***p*-value (Upper Tail)**	0.0062	
15	Male		***p*-value (Two Tail)**	0.0124	
16	Female				
400	Male				
401	Male				
402					

Nota: Los renglones 17 a 399 se encuentran ocultos.

poblacional. Si los datos muestrales ya han sido resumidos, no es necesario ingresarlos en la hoja de cálculo. En este caso, para obtener los resultados se ingresa el tamaño de la muestra en la celda D3, la proporción muestral en la celda D6 y el valor hipotético de la proporción poblacional en la celda D8. La hoja de cálculo que se presenta en la figura 9.13 se encuentra bajo el nombre Hypothesis p en el disco compacto que viene con el libro.

CAPÍTULO 10



Inferencia estadística acerca de medias y de proporciones con dos poblaciones

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

FOOD AND DRUG ADMINISTRATION DE ESTADOS UNIDOS
WASHINGTON, D. C.

La Food and Drug Administration de Estados Unidos (FDA), a través del Center for Drug Evaluation and Research (CDER), garantiza que los medicamentos sean confiables y efectivos. Pero el CDER no se encarga de realizar las pruebas necesarias a los medicamentos nuevos. La empresa interesada en producir un nuevo medicamento es la responsable de presentar las evidencias de que el medicamento es confiable y efectivo. Después, en el CDER, científicos y especialistas en estadística revisan las evidencias presentadas.

Las empresas interesadas en que se apruebe un nuevo medicamento realizan extensos estudios estadísticos para apoyar su solicitud. Las pruebas que se realizan en la industria farmacéutica suelen comprender tres etapas: 1) pruebas preclínicas, 2) pruebas de uso prolongado y confiabilidad y 3) pruebas de eficiencia clínica. En cada una de las etapas sucesivas disminuye la posibilidad de que el medicamento pase las rigurosas pruebas; en cambio, el costo de las pruebas subsiguientes aumenta enormemente. Los estudios realizados informan que el costo promedio de la investigación y desarrollo de un nuevo medicamento es de $250 millones y la duración es 12 años. De esta manera, es importante descartar ya en las pruebas de las primeras etapas aquellos medicamentos que no resultarán útiles e identificar los que parecen prometedores para continuar sometiéndolos a las distintas pruebas.



Los métodos estadísticos se usan para desarrollar y probar medicamentos nuevos. © Lester Lefkowitz/ CORBIS.

La estadística tiene un papel muy importante en la investigación farmacéutica, para la cual existen disposiciones gubernamentales estrictas y rigurosas. En las pruebas preclínicas suelen emplearse pruebas estadísticas en las que intervienen dos o tres poblaciones para determinar si se debe continuar con las pruebas de uso prolongado y confiabilidad del nuevo medicamento. Las poblaciones son: una para el nuevo medicamento, como control, y otra para un medicamento estándar. Los estudios o pruebas preclínicas inician con el envío del medicamento al departamento de farmacología para que evalúen su eficacia y su capacidad para producir los efectos esperados. Como parte de este proceso, se le pide a un especialista en estadística que diseñe un experimento para probar el nuevo medicamento. En este diseño se especifica el tamaño de la muestra y los métodos estadísticos de análisis. En los estudios en los que intervienen dos poblaciones, se usa una muestra para obtener datos sobre la eficacia del nuevo medicamento (población 1) y otra para obtener datos acerca de la eficacia de un medicamento estándar (población 2). Dependiendo del uso que tendrá el nuevo medicamento en disciplinas como la neurología, la cardiología, la inmunología, etc., se probará el nuevo medicamento y el medicamento estándar. En la mayor parte de los estudios se emplea una prueba de hipótesis para determinar la diferencia entre las medias de la población del nuevo medicamento y de la población del medicamento estándar. Si el nuevo medicamento es menos eficaz que el medicamento estándar o tiene efectos indeseables, se rechaza y queda eliminado del programa de pruebas. Sólo el nuevo medicamento que muestra comparaciones prometedoras respecto del medicamento estándar, continúa con las pruebas.

En las etapas de pruebas de uso prolongado y confiabilidad y de eficiencia clínica también se recolectan más datos y se realizan otros estudios multipoblacionales. La FDA requiere que los métodos estadísticos se definan antes de que se realicen las pruebas para evitar sesgos. Además, para impedir sesgos humanos, algunos de los estudios clínicos son doble o triple ciegos. Es decir, ni el paciente ni el investigador saben quién recibe cuál medicamento. Si el nuevo medicamento satisface todos los requerimientos en comparación al medicamento estándar, se presenta una solicitud de aceptación del nuevo medicamento ante la FDA. La solicitud es rigurosamente analizada por los científicos y especialistas en estadística de esta agencia.

En este capítulo se verá cómo calcular intervalos de confianza y realizar pruebas de hipótesis para medias y proporciones cuando se comparan dos poblaciones. Se presentarán las técnicas para analizar tanto muestras aleatorias independientes como muestras pareadas.

En los capítulos 8 y 9 se mostró cómo obtener un intervalo de confianza para realizar una prueba de hipótesis cuando se trata de una sola media poblacional o de una sola proporción poblacional. En este capítulo se continúa con el estudio de la inferencia estadística mostrando la forma de estimar por intervalo y realizar pruebas de hipótesis cuando se tienen dos poblaciones y lo que interesa es la diferencia entre dos medias poblacionales o entre dos proporciones poblacionales. Por ejemplo, quizá desee obtener una estimación por intervalo para la diferencia entre el salario inicial medio de la población de hombres y el salario inicial medio de la población de mujeres, o necesite realizar una prueba de hipótesis para determinar si hay alguna diferencia entre la proporción de piezas defectuosas producidas por el proveedor A y la proporción de partes defectuosas producidas por el proveedor B. El estudio de inferencia estadística para dos poblaciones se inicia mostrando la forma de obtener una estimación mediante un intervalo de confianza y cómo realizar una prueba de hipótesis para la diferencia entre las medias de dos poblaciones en el caso en que se considera que se conocen las desviaciones estándar de estas dos poblaciones.



Inferencias acerca de la diferencia entre dos medias poblacionales: σ_1 y σ_2 conocidas

Sean μ_1 la media de la población 1 y μ_2 la media de la población 2, lo que interesa aquí son inferencias acerca de la diferencia entre las medias: $\mu_1 - \mu_2$. Para hacer una inferencia acerca de esta diferencia, se elige una muestra aleatoria simple de n_1 unidades de la población 1 y otra muestra aleatoria simple de n_2 unidades de la población 2. A estas dos muestras que se toman separada e independientemente se les conoce como **muestras aleatorias simples independientes**. En esta sección se supondrá que se cuenta con información que permite considerar que las dos desviaciones estándar σ_1 y σ_2 se conocen antes de tomar las muestras. Este caso se conoce como el caso σ_1 y σ_2 conocidas. En el ejemplo siguiente, se muestra cómo calcular el margen de error y obtener una estimación por intervalo para la diferencia entre las dos medias poblacionales cuando se conocen σ_1 y σ_2.

Estimación por intervalo de $\mu_1 - \mu_2$

Greystone Department Stores, Inc. tiene dos tiendas en Buffalo, Nueva York, una en el centro de la ciudad y otra en un centro comercial. El gerente regional ha observado que los productos que se venden bien en una tienda no se venden bien en la otra. El gerente cree que esto se debe a diferencias demográficas entre los clientes de las dos tiendas. Debe haber diferencias de edad, educación, ingreso, etc., entre los clientes de una y otra tienda. Suponga que el gerente pide que se investigue la diferencia entre las medias de las edades de los clientes de las dos tiendas.

Si la población 1 es la población de clientes que compra en la tienda del centro de la ciudad y la población 2 es la población de clientes que compra en la tienda del centro comercial, se tiene:

μ_1 = media de la población 1 (es decir, media de las edades de los clientes que compran en la tienda del centro de la ciudad)

μ_2 = media de la población 2 (es decir, media de las edades de los clientes que compran en la tienda del centro comercial).

La diferencia entre las dos medias poblacionales es $\mu_1 - \mu_2$.

Para estimar $\mu_1 - \mu_2$, se toma una muestra aleatoria simple de n_1 clientes de la población 1 y una muestra aleatoria simple de n_2 clientes de la población 2, y se calculan las dos medias muestrales:

$\bar{x}_1$ = media muestral de las edades en la muestra aleatoria simple de n_1 clientes del centro de la ciudad

$\bar{x}_2$ = media muestral de las edades en la muestra aleatoria simple de n_2 clientes del centro comercial

La estimación puntual de la diferencia entre las dos medias poblacionales es la diferencia entre las dos medias muestrales.

ESTIMADOR PUNTUAL DE LA DIFERENCIA ENTRE LAS DOS MEDIAS POBLACIONALES

$$\bar{x}_1 - \bar{x}_2 \tag{10.1}$$

En la figura 10.1 se proporciona una visión esquemática del proceso que se sigue para estimar la diferencia entre dos medias poblacionales empleando dos muestras aleatorias simples.

El error estándar de $\bar{x}_1 - \bar{x}_2$ es la desviación estándar de la distribución muestral de $\bar{x}_1 - \bar{x}_2$.

Como otros estimadores puntuales, el estimador puntual $\bar{x}_1 - \bar{x}_2$ tiene un error estándar que describe la variación de la distribución muestral del estimador. Cuando se tienen dos muestras aleatorias simples independientes, el error estándar de $\bar{x}_1 - \bar{x}_2$ es el siguiente:

ERROR ESTÁNDAR DE $\bar{x}_1 - \bar{x}_2$

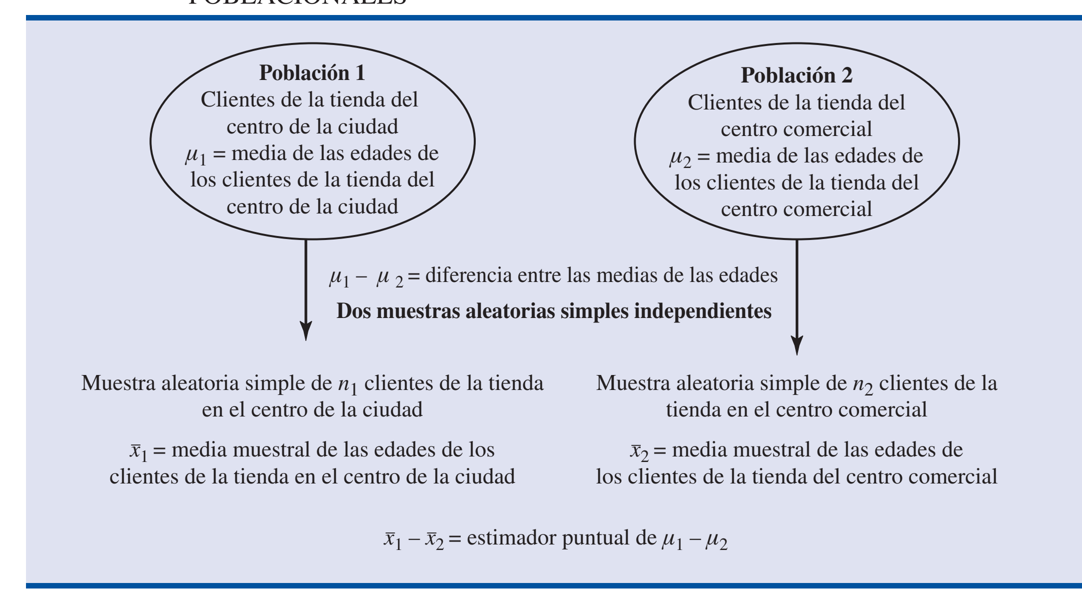
$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{10.2}$$

Si ambas poblaciones tienen distribución normal o si los tamaños de las muestras son suficientemente grandes para que el teorema del límite central permita concluir que las distribuciones muestrales de $\bar{x}_1$ y $\bar{x}_2$ puedan ser aproximadas mediante una distribución normal, la distribución muestral de $\bar{x}_1 - \bar{x}_2$ tendrá una distribución normal cuya media es $\mu_1 - \mu_2$.

Como se mostró en el capítulo 8, una estimación por intervalo está dada por una estimación puntual ± un margen de error. En el caso de la estimación de la diferencia entre dos medias poblacionales, una estimación por intervalo tendrá la forma siguiente:

$$\bar{x}_1 - \bar{x}_2 \pm \text{Margen de error}$$

FIGURA 10.1 ESTIMACIÓN DE LA DIFERENCIA ENTRE DOS MEDIAS POBLACIONALES



Como la distribución muestral de $\bar{x}_1 - \bar{x}_2$ tiene una distribución normal, el margen de error se expresa de la manera siguiente:

El margen de error se obtiene multiplicando el error estándar por $z_{\alpha/2}$.

$$\text{Margen de error} = z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2} = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \textbf{(10.3)}$$

En consecuencia, la estimación por intervalo de la diferencia entre las dos medias poblacionales es la siguiente:

ESTIMACIÓN POR INTERVALO DE LA DIFERENCIA ENTRE DOS MEDIAS POBLACIONALES: σ_1 Y σ_2 CONOCIDAS

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \textbf{(10.4)}$$

donde $1 - \alpha$ es el coeficiente de confianza.

De regreso al ejemplo de la tienda de departamentos Greystone: de acuerdo con datos de anteriores estudios demográficos sobre los clientes, las dos desviaciones estándar poblaciones se conocen y son $\sigma_1 = 9$ años y $\sigma_2 = 10$ años. De los datos de las dos muestras aleatorias simples independientes de los clientes de Greystone se obtuvieron los resultados siguientes.

	Tienda en el centro de la ciudad	**Tienda en el centro comercial**
Tamaño de la muestra	$n_1 = 36$	$n_2 = 49$
Media muestral	$\bar{x}_1 = 40$ años	$\bar{x}_2 = 35$ años

Mediante la expresión 10.1 se encuentra que la estimación puntual de la diferencia entre las dos medias poblacionales es $\bar{x}_1 - \bar{x}_2 = 40 - 35 = 5$ años. Por ende, se estima que, en promedio, los clientes de la tienda en el centro de la ciudad son cinco años mayores que los clientes de la tienda del centro comercial. Ahora, con la expresión 10.4 se calcula el margen de error y se proporciona una estimación por intervalo de $\mu_1 - \mu_2$. Si tiene 95% de confianza y $z_{\alpha/2} = z_{0.025} = 1.96$:

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$40 - 35 \pm 1.96\sqrt{\frac{9^2}{36} + \frac{10^2}{49}}$$

$$5 \pm 4.06$$

El margen de error es 4.06 años y la estimación por intervalo de 95% de confianza de la diferencia entre las medias poblacionales va de $5 - 4.06 = 0.94$ años a $5 + 4.06 = 9.06$ años.

Prueba de hipótesis acerca de $\mu_1 - \mu_2$

Ahora se verán las pruebas de hipótesis acerca de la diferencia entre dos medias poblacionales. D_0 denota la diferencia hipotética entre μ_1 y μ_2, las tres formas que puede tener una prueba de hipótesis son las siguientes:

$$H_0\colon \mu_1 - \mu_2 \geq D_0 \qquad H_0\colon \mu_1 - \mu_2 \leq D_0 \qquad H_0\colon \mu_1 - \mu_2 = D_0$$
$$H_a\colon \mu_1 - \mu_2 < D_0 \qquad H_a\colon \mu_1 - \mu_2 > D_0 \qquad H_a\colon \mu_1 - \mu_2 \neq D_0$$

En muchas aplicaciones $D_0 = 0$. Con un ejemplo de una prueba de hipótesis de dos colas, cuando $D_0 = 0$ la hipótesis nula es H_0: $\mu_1 - \mu_2 = 0$. En este caso, la hipótesis nula es que μ_1 y μ_2 son iguales. Rechazar H_0 lleva a la conclusión de que H_a: $\mu_1 - \mu_2 \neq 0$ es verdadera: μ_1 y μ_2 no son iguales.

Los pasos presentados en el capítulo 9 para realizar una prueba de hipótesis también son aplicables aquí. Hay que elegir el nivel de significancia, calcular el valor del estadístico de prueba y encontrar el valor-*p* para determinar si se rechaza la hipótesis nula. En el caso de dos muestras aleatorias independientes, se mostró que el estimador puntual $\bar{x}_1 - \bar{x}_2$ tiene un error estándar $\sigma_{\bar{x}_1 - \bar{x}_2}$ dado por la expresión (10.2) y, cuando los tamaños de las muestras son suficientemente grandes, la distribución de $\bar{x}_1 - \bar{x}_2$ se puede considerar como una distribución normal. En este caso, el estadístico de prueba para la diferencia entre dos medias poblacionales cuando se conocen σ_1 y σ_2 es el que se da a continuación.

ESTADÍSTICO DE PRUEBA PARA PRUEBA DE HIPÓTESIS ACERCA DE $\mu_1 - \mu_2$: σ_1 Y σ_2 CONOCIDAS

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad \textbf{(10.5)}$$

El uso de dicho estadístico de prueba se muestra en el siguiente ejemplo de prueba de hipótesis.

Como parte de un estudio para evaluar las diferencias en la calidad entre dos centros de enseñanza, se aplica un examen estandarizado a los individuos de ambos centros. La diferencia de calidad se evalúa comparando las medias de las puntuaciones obtenidas en el examen. Las medias poblacionales en cada uno de los centros son:

μ_1 = media de las puntuaciones en los exámenes presentados por los individuos del centro A

μ_2 = media de las puntuaciones en los exámenes presentados por los individuos del centro B

Debe partir de la suposición tentativa de que no hay diferencia entre la calidad de la capacitación en uno y otro centro de enseñanza. Entonces, en términos de las puntuaciones medias obtenidas en el examen, la hipótesis nula es que $\mu_1 - \mu_2 = 0$. Si las evidencias muestrales llevan a rechazar esta hipótesis, se concluirá que sí hay diferencia entre las medias de las puntuaciones de examen en las dos poblaciones. Esta conclusión indicará que hay diferencia en la calidad de los dos centros y sugerirá la necesidad de realizar un estudio para investigar las razones de estas diferencias. Las hipótesis nula y alternativa en esta prueba de dos colas se expresan como se indica a continuación.

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

En exámenes estandarizados practicados ya en diversas ocasiones, siempre se ha obtenido una desviación estándar cercana a 10 puntos. Por tanto, usará esta información y considerará que las desviaciones estándar poblacionales se conocen y que son $\sigma_1 = 10$ y $\sigma_2 = 10$. Para este estudio se especifica un nivel de significancia $\alpha = 0.05$.



Con muestras aleatorias simples independientes de $n_1 = 30$ individuos del centro de enseñanza A y $n_2 = 40$ individuos del centro de enseñanza B. Las medias muestrales correspondientes son $\bar{x}_1 = 82$ y $\bar{x}_2 = 78$. ¿Estos datos indican que existe una diferencia significativa entre las

medias poblacionales de los dos centros de enseñanza? Para responder esta pregunta se calcula el estadístico de prueba empleando la ecuación (10.5).

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{(82 - 78) - 0}{\sqrt{\dfrac{10^2}{30} + \dfrac{10^2}{40}}} = 1.66$$

A continuación se calcula el valor-p de esta prueba de dos colas. Como el estadístico de prueba z se encuentra en la cola superior, se calcula primero el área bajo la curva a la derecha de $z = 1.66$. En la tabla de la distribución normal estándar, el área a la izquierda de $z = 1.66$ es 0.9515. Por ende, el área en la cola superior de la distribución es $1.000 - 0.9515 = 0.0485$. Como es una prueba de dos colas, hay que duplicar el área que queda a la cola: el valor-$p = 2(0.0485) = 0.0970$. Como la regla es rechazar H_0 si el valor-$p \leq \alpha$, puesto que el valor-p es 0.0970, no se puede rechazar H_0 al nivel de significancia 0.05. Los resultados muestrales no proporcionan suficiente evidencia para concluir que hay una diferencia de calidad entre los dos centros de enseñanza.

En este capítulo, para las pruebas de hipótesis, se usará el método del valor-p descrito en el capítulo 9. Sin embargo, si el estudiante así lo prefiere, use el estadístico de prueba y la regla de rechazo del valor crítico. Para $\alpha = 0.05$ y $z_{\alpha/2} = z_{0.025} = 1.96$, la regla de rechazo empleando el método del valor crítico será, rechazar H_0 si $z \leq 1.96$ o si $z \leq -1.96$. Como $z = 1.66$, se llega a la misma conclusión de no rechazar H_0.

En los siguientes ejemplos se demostrará la prueba de hipótesis de dos colas para encontrar la diferencia entre dos medias poblacionales. Las pruebas de la cola superior e inferior también se considerarán. Dichas pruebas aplican el mismo estadístico de prueba dado en la ecuación (10.5). El procedimiento para calcular el valor-p y las reglas de rechazo para pruebas de una cola son las que se presentaron en el capítulo 9.

Recomendación práctica

En la mayor parte de las aplicaciones de intervalos de estimaciones y de pruebas de hipótesis presentados en esta sección, se consideran adecuadas muestras aleatorias $n_1 \geq 30$ y $n_2 \geq 30$. En el caso en que una o las dos muestras sea menor que 30, las distribuciones de las poblaciones son importantes. En general, cuando las muestras son pequeñas, es importante que el analista se convenza de que es razonable suponer que las distribuciones de las dos poblaciones son por lo menos aproximadamente normales.

Ejercicios

Métodos

1. Los resultados siguientes provienen de muestras aleatorias simples independientes tomadas de dos poblaciones



Muestra 1	Muestra 2
$n_1 = 50$	$n_2 = 35$
$\bar{x}_1 = 13.6$	$\bar{x}_2 = 11.6$
$\sigma_1 = 2.2$	$\sigma_2 = 3.0$

a. ¿Cuál es la estimación puntual de la diferencia entre las dos medias poblacionales?
b. Dé un intervalo de confianza de 90% para la diferencia entre las dos medias poblaciones.
c. Proporcione un intervalo de confianza de 95% para la diferencia entre las dos medias poblaciones.



2. Considere la prueba de hipótesis que se da a continuación.

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_a: \mu_1 - \mu_2 > 0$$

Los resultados siguientes son los resultados obtenidos de dos muestras independientes tomadas de dos poblaciones.

Muestra 1	Muestra 2
$n_1 = 40$	$n_2 = 50$
$\bar{x}_1 = 25.2$	$\bar{x}_2 = 22.8$
$\sigma_1 = 5.2$	$\sigma_2 = 6.0$

a. ¿Cuál es el valor del estadístico de prueba?
b. ¿Cuál es el valor-p?
c. Si $\alpha = 0.05$, ¿cuál es la conclusión de la prueba de hipótesis?

3. Considere la prueba de hipótesis que se da a continuación

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

Los siguientes son los resultados de dos muestras independientes tomadas de dos poblaciones.

Muestra 1	Muestra 2
$n_1 = 80$	$n_2 = 70$
$\bar{x}_1 = 104$	$\bar{x}_2 = 106$
$\sigma_1 = 8.4$	$\sigma_2 = 7.6$

a. ¿Cuál es el valor del estadístico de prueba?
b. ¿Cuál es el valor-p?
c. Si $\alpha = 0.05$, ¿cuál es la conclusión de la prueba de hipótesis?

Aplicaciones

4. Durante el 2003 los precios de la gasolina alcanzaron record de precios altos en 16 estados de Estados Unidos (*The Wall Street Journal*, 7 de marzo de 2003). Dos de los estados afectados fueron California y Florida. La American Automobile Association encontró como precio medio muestral por galón $2.04 en California y $1.72 por galón en Florida. Use 40 como tamaño de la muestra de California y 35 como tamaño de la muestra en Florida. Suponga que estudios anteriores indican que la desviación estándar poblacional en California es 0.10 y en Florida 0.08.
   a. ¿Cuál es la estimación puntual de la diferencia entre los precios medios poblacionales por galón en California y Florida?
   b. ¿Cuál es el margen de error con un 95% de confianza?
   c. ¿Cuál es la estimación por intervalo de 95% de confianza para la diferencia entre los precios medios poblacionales por galón en California y en Florida?

5. Se esperaba que el día de San Valentín el desembolso promedio fuera $100.89 (*USA Today*, 13 de febrero de 2006). ¿Hay diferencia en las cantidades que gastan los hombres y las mujeres? El gasto promedio en una muestra de 40 hombres fue de $135.67 y el gasto promedio en una muestra de 30 mujeres fue de $68.64. Por estudios anteriores se sabe que la desviación estándar poblacional en el gasto de los hombres es de $35 y en el gasto de las mujeres es de $20.

a. ¿Cuál es la estimación puntual de la diferencia entre el gasto medio poblacional de los hombres y el gasto medio poblacional de las mujeres?
b. Con 99% de confianza, ¿cuál es el margen de error?
c. Elabore un intervalo de confianza de 99% para la diferencia entre las dos medias poblacionales.

6. Las más de 40 000 empresas de corretaje hipotecario que hay en Estados Unidos, son uno de los negocios más rentables en ese país. Estas empresas de perfil bajo buscan préstamos para sus clientes a cambio de comisiones. La Mortgage Bankers Association of America proporciona datos sobre la magnitud promedio de los préstamos que consiguen estas empresas (*The Wall Street Journal*, 24 de febrero de 2003). El archivo titulado Mortgage del disco compacto contiene los datos de una muestra de 250 préstamos hechos en 2001 y una muestra de 270 préstamos hechos en 2002. De acuerdo con datos anteriores, las desviaciones estándar poblacionales de los montos de los préstamos son \$55 000 en el 2002 y \$50 000 en el 2001. ¿Estos datos muestrales indican que entre 2001 y 2002 ha habido un incremento en el monto medio de los préstamos? Use $\alpha = 0.05$.



Mortgage

7. Durante la temporada de 2003, la Liga Mayor de Béisbol tomó medidas para acelerar el juego en los partidos con objeto de mantener el interés de los aficionados (*CNN Headline News*, 30 de septiembre de 2003). Los resultados siguientes se obtuvieron de una muestra de 60 partidos jugados en el verano de 2002 y de una muestra de 50 partidos jugados en el verano de 2003. La media muestral da la duración media de los juegos que formaron parte de la muestra.

Temporada 2002	Temporada 2003
$n_1 = 60$	$n_2 = 50$
$\bar{x}_1 = 2$ horas, 52 minutos	$\bar{x}_2 = 2$ horas, 46 minutos

a. La hipótesis de investigación era que las medidas tomadas en la temporada de 2003 reducirían la duración media poblacional de los juegos de béisbol. Formule las hipótesis nula y alternativa.
b. ¿Cuál es la estimación puntual de la reducción de la media de duración de los juegos en 2003?
c. Datos de estudios anteriores indican que, para ambos años, la desviación estándar poblacional fue de 12 minutos. Realice la prueba de hipótesis y dé el valor-*p*. Usando como nivel de significancia 0.05, ¿cuál es su conclusión?
d. Dé una estimación por intervalo de 95% de confianza de la duración media de los partidos en el 2003.
e. ¿Cuál es la reducción porcentual en la duración media de los partidos de béisbol en la temporada de 2003? ¿Estarán satisfechos los directivos con los resultados del análisis estadístico? Analice. En los años venideros ¿seguirá siendo un problema la duración de los juegos de béisbol? Explique.

8. Arnold Palmer y Tiger Woods son dos de los mejores golfistas de todos los tiempos. Para comparar a estos dos golfistas en los datos muestrales siguientes se proporcionan los resultados de puntuaciones del hoyo 18 durante un torneo de la PGA. Las puntuaciones de Palmer son de la temporada de 1960 y las de Woods son de la temporada de 1999 (*Golf Magazine*, febrero de 2000).

Palmer, 1960	Woods, 1999
$n_1 = 112$	$n_2 = 84$
$\bar{x}_1 = 69.95$	$\bar{x}_2 = 69.56$

Use los resultados muestrales para probar la hipótesis de que entre los dos jugadores no hay diferencia en las medias poblacionales de las puntuaciones del hoyo 18.

a. Con una desviación estándar poblacional de 2.5 para ambos golfistas, ¿cuál es el valor del estadístico de prueba?
b. ¿Cuál es el valor-*p*?
Si $\alpha = 0.01$, ¿cuál es su conclusión?



10.2 Inferencias acerca de la diferencia entre dos medias poblacionales: σ_1 y σ_2 desconocidas

En esta sección el estudio de las inferencias sobre la diferencia entre dos medias poblacionales se extiende al caso en el que las dos desviaciones estándar poblacionales, σ_1 y σ_2 no se conocen. En este caso, para estimar las desviaciones estándar poblacionales desconocidas se emplean las desviaciones estándar muestrales, s_1 y s_2. Cuando se usan las desviaciones estándar muestrales en las estimaciones por intervalo y en las pruebas de hipótesis, se emplea la distribución t en lugar de la distribución normal estándar.

Estimación por intervalo para $\mu_1 - \mu_2$

En el ejemplo siguiente se muestra cómo calcular el margen de error y obtener una estimación por intervalo para la diferencia entre dos medias poblacionales cuando σ_1 y σ_2 no se conocen. Clearwater National Bank realiza un estudio para identificar diferencias entre las cuentas de cheques de sus clientes en dos de sus sucursales; toma una muestra aleatoria simple de 28 cuentas de la sucursal Cherry Grove y otra muestra aleatoria simple e independiente de 22 cuentas de cheques de la sucursal Beechmont. El saldo se registra en las cuentas de cheques. A continuación se presenta un resumen de los saldos en estas cuentas de cheques.



	Cherry Grove	Beechmont
Tamaño de la muestra	$n_1 = 28$	$n_2 = 22$
Media muestral	$\bar{x}_1 = \$1025$	$\bar{x}_2 = \$910$
Desviación estándar muestral	$s_1 = \$150$	$s_2 = \$125$

El banco desea estimar la diferencia entre el saldo medio en las cuentas de cheques de la población de clientes de Cherry Grove y el saldo medio en las cuentas de cheques de la población de clientes de Becchmont. A continuación se calculará el margen de error y se dará una estimación por intervalo para la diferencia entre estas dos medias poblacionales.

En la sección 10.1 se proporcionó la estimación por intervalo siguiente para el caso en el que se conocen las dos desviaciones estándar poblacionales σ_1 y σ_2.

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Cuando σ_1 y σ_2 se estiman mediante s_1 y s_2, se usa la distribución t para hacer inferencias sobre la diferencia entre dos medias poblacionales.

Cuando no se conocen σ_1 y σ_2 se emplean s_1 y s_2 para estimar σ_1 y σ_2 y $z_{\alpha/2}$ se sustituye por $t_{\alpha/2}$. Entonces, la estimación por intervalo para la diferencia entre dos medias poblacionales queda dada por la expresión siguiente:

ESTIMACIÓN POR INTERVALO PARA LA DIFERENCIA ENTRE DOS MEDIAS POBLACIONALES: σ_1 Y σ_2 DESCONOCIDAS

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \textbf{(10.6)}$$

donde $1 - \alpha$ es el coeficiente de confianza.

En esta expresión el uso de la distribución t es una aproximación, pero proporciona resultados excelentes y es relativamente fácil de usar. La única dificultad que se encuentra al emplear la expresión (10.6) es determinar los grados de libertad para $t_{\alpha/2}$. Los paquetes de software calculan automáticamente los grados de libertad. La fórmula que se usa es la siguiente:

GRADOS DE LIBERTAD: DISTRIBUCIÓN t CON DOS MUESTRAS ALEATORIAS INDEPENDIENTES

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} \tag{10.7}$$

Al retomar el ejemplo del banco se mostrará cómo usar la expresión (10.6) para obtener un intervalo de confianza de 95% para estimar la diferencia entre las medias poblacionales de los saldos en las cuentas de cheques en sus dos sucursales. Los datos muestrales de la sucursal Cherry Grove son $n_1 = 28$, $\bar{x}_1 = \$1025$, $s_1 = \$150$ y los de la sucursal Beechmont son $n_2 = 22$, $\bar{x}_2 = \$910$, $s_2 = \$125$. El cálculo de los grados de libertad para $t_{\alpha/2}$ es:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{150^2}{28} + \frac{125^2}{22}\right)^2}{\frac{1}{28 - 1}\left(\frac{150^2}{28}\right)^2 + \frac{1}{22 - 1}\left(\frac{125^2}{22}\right)^2} = 47.8$$

Como el resultado no es un número entero, se redondea hacia *abajo* a 47 para tener un valor t mayor y dar una estimación por intervalo más prudente. En la tabla de la distribución t para 47 grados de libertad, se encuentra $t_{0.025} = 2.012$. De acuerdo con la expresión (10.6), el intervalo de confianza de 95% para la diferencia entre las dos medias poblacionales se calcula como sigue.

$$\bar{x}_1 - \bar{x}_2 \pm t_{.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$1025 - 910 \pm 2.012\sqrt{\frac{150^2}{28} + \frac{125^2}{22}}$$

$$115 \pm 78$$

La estimación puntual de la diferencia entre las dos medias poblacionales de los saldos en las cuentas de cheques es \$115. El margen de error es \$78 y la estimación por intervalo de 95% de confianza para la diferencia entre las dos medias poblacionales es el que va de $115 - 78 = \$37$ a $115 + 78 = \$193$.

Esta sugerencia es útil cuando se usa la ecuación (10.7) para calcular a mano los grados de libertad.

El cálculo a mano de los grados de libertad usando la expresión (10.7) es laborioso, pero muy sencillo si se usa un paquete de software. Sin embargo, observe que las expresiones s_1^2/n_1 y s_2^2/n_2 aparecen tanto en la expresión (10.6) como en la expresión (10.7). Por tanto, sólo habrá que calcular estas expresiones una vez para usarlas en ambas expresiones, (10.6) y (10.7).

Pruebas de hipótesis acerca de $\mu_1 - \mu_2$

Ahora se estudiarán las pruebas de hipótesis acerca de la diferencia entre las medias de dos poblaciones cuando no se conocen las desviaciones estándar poblacionales σ_1 y σ_2. Sea D_0 la dife-

rencia hipotética entre μ_1 y μ_2. en la sección 10.1 se mostró que el estadístico de prueba usado cuando se conocen σ_1 y σ_2 es el siguiente:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

El estadístico de prueba z sigue la distribución normal estándar.

Cuando no se conocen σ_1 y σ_2, se usa s_1 para estimar σ_1 y s_2 para estimar s_2. Sustituyendo σ_1 y σ_2 por estas desviaciones estándar muestrales se obtiene el siguiente estadístico de prueba para el caso en el que no se conocen σ_1 y σ_2.

ESTADÍSTICO DE PRUEBA PARA PRUEBAS DE HIPÓTESIS ACERCA DE μ_1 Y μ_2: σ_1 Y σ_2 DESCONOCIDAS

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{10.8}$$

Los grados de libertad para la distribución t se obtienen con la ecuación (10.7).

Ahora se demostrará el uso del estadístico de prueba en el ejemplo siguiente.

Considere un nuevo software que ayuda a los analistas de sistemas a reducir el tiempo requerido para diseñar, elaborar y poner en marcha un sistema de información. Para evaluar las ventajas del nuevo software, se toma una muestra de 24 analistas de sistemas. A cada analista se le da información sobre un sistema de información hipotético. A 12 de ellos se les pide que elaboren el sistema de información usando la tecnología existente y a los otros 12 analistas se les capacita para usar el nuevo software y se les pide que lo empleen para elaborar el sistema de información.

En el estudio participan dos poblaciones: una población de analistas de sistema que usan la tecnología ya existente y una población de analistas de sistemas que usan el nuevo software. En términos del tiempo necesario para el proyecto del sistema de información, las medias poblacionales son las siguientes:

μ_1 = media del tiempo que necesitan para el proyecto los analistas que emplean la tecnología ya existente.

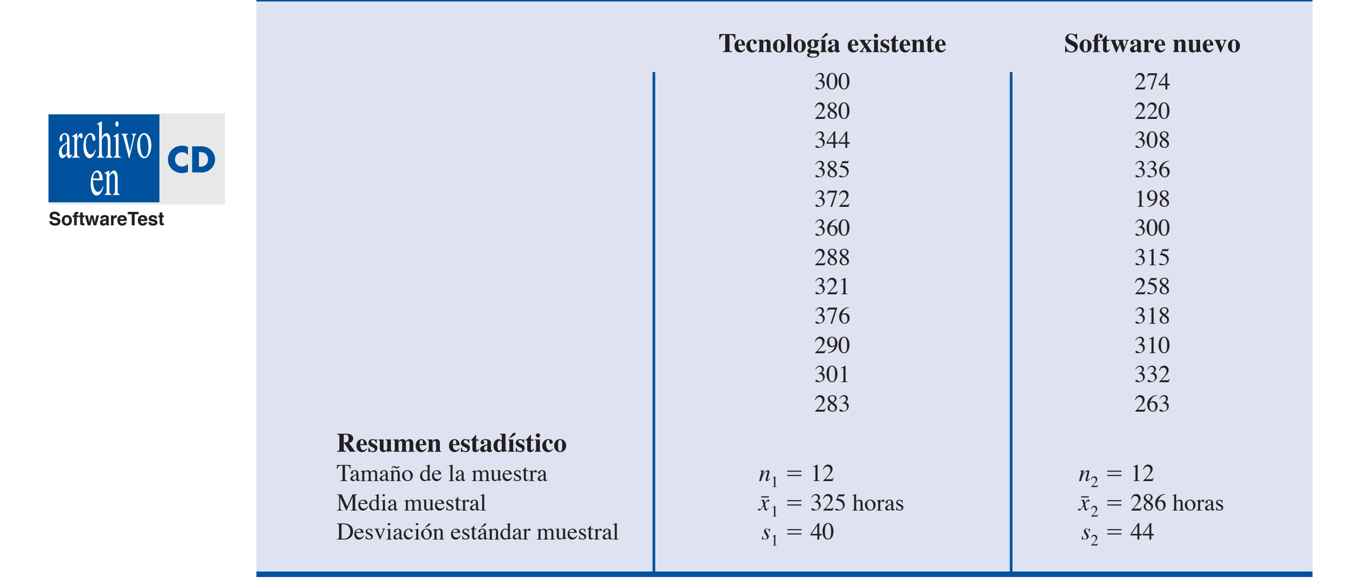
μ_2 = media del tiempo que necesitan para el proyecto los analistas que emplean el nuevo software.

El investigador encargado de la evaluación del nuevo software espera poder demostrar que con el nuevo software se necesita menos tiempo para el proyecto del sistema de información. De manera que el investigador tratará de hallar evidencias que le permitan concluir que μ_2 es menor que μ_1, caso en el que la diferencia $\mu_1 - \mu_2$ será mayor que cero. La hipótesis de investigación $\mu_1 - \mu_2 > 0$ se establece como la hipótesis alternativa. Por lo que la prueba de hipótesis será

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_a: \mu_1 - \mu_2 > 0$$

Como nivel de significancia se usará $\alpha = 0.05$.

TABLA 10.1 DATOS Y RESUMEN ESTADÍSTICO DEL TIEMPO REQUERIDO EN EL ESTUDIO DE LA PRUEBA DE SOFTWARE



	Tecnología existente	Software nuevo
	300	274
	280	220
	344	308
	385	336
	372	198
	360	300
	288	315
	321	258
	376	318
	290	310
	301	332
	283	263
Resumen estadístico		
Tamaño de la muestra	$n_1 = 12$	$n_2 = 12$
Media muestral	$\bar{x}_1 = 325$ horas	$\bar{x}_2 = 286$ horas
Desviación estándar muestral	$s_1 = 40$	$s_2 = 44$

Suponga que los resultados de los 24 analistas son los que se presentan en la tabla 10.1. Con el estadístico de prueba dado en la ecuación (10.8) se tiene,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{(325 - 286) - 0}{\sqrt{\dfrac{40^2}{12} + \dfrac{44^2}{12}}} = 2.27$$

De acuerdo con la ecuación (10.7) los grados de libertad son

$$gl = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{1}{n_1 - 1}\left(\dfrac{s_1^2}{n_1}\right)^2 + \dfrac{1}{n_2 - 1}\left(\dfrac{s_2^2}{n_2}\right)^2} = \frac{\left(\dfrac{40^2}{12} + \dfrac{44^2}{12}\right)^2}{\dfrac{1}{12 - 1}\left(\dfrac{40^2}{12}\right)^2 + \dfrac{1}{12 - 1}\left(\dfrac{44^2}{12}\right)^2} = 21.8$$

Al redondear hacia abajo, se usará una distribución t con 21 grados de libertad. El renglón correspondiente de la distribución t es:

Área en la cola superior	0.20	0.10	0.05	0.025	0.01	0.005
Valor t (21 gl)	0.859	1.323	1.721	2.080	2.518	2.831

$t = 2.27$

Mediante la tabla de la distribución t, sólo se puede determinar un rango para el valor-p. Si se usa Excel o Minitab, se obtiene exactamente el valor-p = 0.017.

En una prueba de la cola superior, el valor-p es el área en la cola superior a la derecha de $t = 2.27$. De acuerdo con este resultado se ve que el valor-p está entre 0.025 y 0.01. Por tanto, el valor-p es menor que $\alpha = 0.05$ y se rechaza H_0. Los resultados muestrales permiten al investigador concluir que $\mu_1 - \mu_2 > 0$, o que $\mu_1 > \mu_2$. La investigación favorece la conclusión de que con el nuevo software el tiempo requerido es menor.

FIGURA 10.2 RESULTADO PROPORCIONADO CON MINITAB PARA LA PRUEBA DE HIPÓTESIS USANDO LA TECNOLOGÍA EXISTENTE Y EL NUEVO SOFTWARE

```
Two-sample T for Current vs New

            N     Mean     StDev    SE Mean
Current    12    325.0      40.0         12
New        12    286.0      44.0         13

Difference = mu Current - mu New
Estimate for difference:  39.0000
95% lower bound for difference = 9.4643
T-Test of difference = 0 (vs >):  T-Value = 2.27  P-Value = 0.017  DF = 21
```

Para las pruebas de hipótesis acerca de la diferencia entre dos medias poblacionales se usan Excel o Minitab. En la tabla 10.1 se presentan los resultados que proporciona Minitab en esta comparación de la tecnología existente y el nuevo software. En la última línea se ve que $t = 2.27$ y que el valor-$p = 0.017$. Observe que Minitab usa la ecuación (10.7) para calcular los 21 grados de libertad.

Recomendación práctica

Se recomienda, siempre que sea posible, usar muestras del mismo tamaño, $n_1 = n_2$.

Los procedimientos aquí presentados para estimaciones por intervalo y para pruebas de hipótesis son robustos y pueden usarse con muestras relativamente pequeñas. En la mayor parte de las aplicaciones con muestras casi del mismo tamaño y de manera que el tamaño total de la muestra, $n_1 + n_2$, sea por lo menos 20 se esperan muy buenos resultados, aun cuando las poblaciones no sean normales. Si las distribuciones de las poblaciones son muy sesgadas o contienen valores atípicos se recomienda usar muestras más grandes. Muestras pequeñas sólo deben usarse cuando el analista está convencido de que las distribuciones de las poblaciones con aproximadamente normales.

NOTAS Y COMENTARIOS

Otro método que se usa para hacer inferencias acerca de la diferencia entre dos medias poblacionales cuando no se conocen σ_1 y σ_2 se basa en la suposición de que las dos desviaciones estándar son *iguales* ($\sigma_1 = \sigma_2 = \sigma$). Cuando se usa esta suposición, las dos desviaciones estándar muestrales se combinan para obtener la siguiente *varianza muestral combinada:*

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

El estadístico de prueba es

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

el cual tiene $n_1 + n_2 - 2$ grados de libertad. A partir de aquí, el cálculo del valor-p y la interpretación de los resultados muestrales se hacen de manera idéntica a lo indicado en esta sección.

El problema con este procedimiento es la dificultad que suele existir para verificar que las dos desviaciones estándar poblacionales son iguales. Lo más frecuente es encontrar desviaciones estándar poblacionales que no son iguales. Con el procedimiento de la varianza combinada pueden no obtenerse resultados satisfactorios, en especial si los tamaños de las muestras, n_1 y n_2, son muy distintos.

El procedimiento presentado en esta sección no requiere la suposición de que las dos desviaciones estándar poblacionales son iguales y puede usarse cuando las dos desviaciones estándar poblacionales son o no iguales. Es un procedimiento más general y se recomienda para la mayoría de las aplicaciones.

Ejercicios

Métodos



9. Los resultados siguientes se obtuvieron de muestras aleatorias independientes tomadas de dos poblaciones.

Muestra 1	Muestra 2
$n_1 = 20$	$n_2 = 30$
$\bar{x}_1 = 22.5$	$\bar{x}_2 = 20.1$
$s_1 = 2.5$	$s_2 = 4.8$

a. ¿Cuál es la estimación puntual de la diferencia entre las dos medias poblacionales?
b. Dé los grados de libertad para la distribución *t*.
c. Con 95% de confianza, ¿cuál es el margen de error?
d. Dé el intervalo de 95% de confianza para la diferencia entre las dos medias poblacionales.

Autoexamen

10. Considere la prueba de hipótesis siguiente

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

Los resultados siguientes se obtuvieron de muestras independientes tomadas de dos poblaciones.

Muestra 1	Muestra 2
$n_1 = 35$	$n_2 = 40$
$\bar{x}_1 = 13.6$	$\bar{x}_2 = 10.1$
$s_1 = 5.2$	$s_2 = 8.5$

a. ¿Cuál es el valor del estadístico de prueba?
b. Dé los grados de libertad para la distribución *t*.
c. ¿Cuál es el valor-*p*?
d. Con $\alpha = 0.05$, ¿cuál es la conclusión?

11. Los datos siguientes se obtuvieron de dos muestras aleatorias independientes tomadas de dos poblaciones.

Muestra 1	10	7	13	7	9	8
Muestra 2	8	7	8	4	6	9

a. Calcule las dos medias muestrales.
b. Calcule las dos desviaciones estándar muestrales.
c. Dé la estimación puntual de la diferencia entre las dos medias poblacionales.
d. Dé la estimación por intervalo de 95% de confianza para la diferencia entre las dos medias poblacionales.

Aplicaciones



12. El U.S. Department of Transportation informa sobre la cantidad de millas que recorren en automóvil los habitantes de las 75 principales áreas metropolitanas de ese país. Suponga que en una muestra aleatoria simple de 50 habitantes de Buffalo, la media es 22.5 millas por día y la desvia-

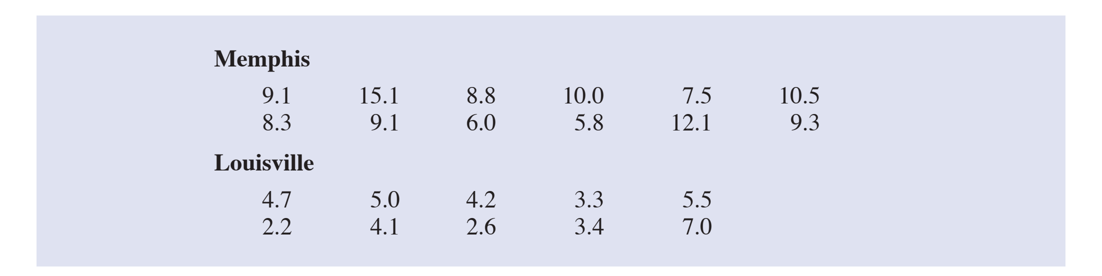
ción estándar es 8.4 millas por día y que en una muestra aleatoria simple independiente de 40 habitantes de Boston la media es 18.6 millas por día y la desviación estándar es 7.4 millas por día.

a. ¿Cuál es la estimación puntual de la diferencia entre la media de las millas por día que recorre un habitante de Buffalo y la media de las millas por día que recorre un habitante de Boston?
b. Dé un intervalo de confianza de 95% para la diferencia entre las dos medias poblacionales.



Cargo

13. FedEx y United Parcel Service (UPS) son las dos empresas de transporte de paquetería más importantes del mundo en cuanto a volumen e ingresos (*The Wall Street Journal*, 27 de enero de 2004). De acuerdo con el Consejo Internacional de Aeropuertos, el aeropuerto internacional de Memphis (FedEx) y el aeropuerto internacional de Louisville (UPS) son dos de los 10 mayores aeropuertos de carga del mundo. Las muestras aleatorias siguientes muestran las toneladas de carga por día que pasan por estos aeropuertos. Los datos están dados en miles de toneladas.



Memphis					
9.1	15.1	8.8	10.0	7.5	10.5
8.3	9.1	6.0	5.8	12.1	9.3
Louisville					
4.7	5.0	4.2	3.3	5.5	
2.2	4.1	2.6	3.4	7.0	

a. Calcule la media muestral y la desviación estándar muestral para cada uno de los aeropuertos.
b. Dé la estimación puntual de la diferencia entre las dos medias poblacionales. Interprete este valor en términos del aeropuerto de mayor volumen y de la diferencia de volúmenes entre los dos aeropuertos.
c. Proporcione un intervalo de 95% de confianza para la diferencia entre las medias poblacionales diarias de los dos aeropuertos.

14. En las zonas costeras de Estados Unidos, Cape Cod, Outer Banks, las Carolinas y la costa del Golfo, hubo, durante los años noventa, un crecimiento relativamente rápido de la población. Los datos recolectados son sobre las personas que viven tanto en zonas costeras como en zonas no costeras de todo Estados Unidos (*USA Today*, 21 de julio de 2000). Suponga que se obtuvieron los resultados muestrales siguientes sobre las edades de estas dos poblaciones de personas.

Zona costera	Zona no costera
$n_1 = 150$	$n_2 = 175$
$\bar{x}_1 = 39.3$ años	$\bar{x}_2 = 35.4$ años
$s_1 = 16.8$ años	$s_2 = 15.2$ años

Pruebe la hipótesis de que no hay diferencia entre las dos medias poblacionales. Use $\alpha = 0.05$.

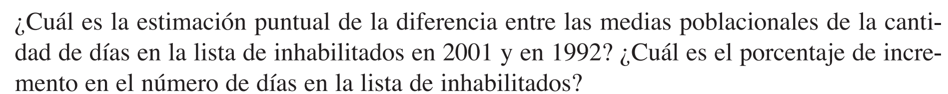
a. Formule las hipótesis nula y alternativa.
b. ¿Cuál es el valor del estadístico de prueba?
c. ¿Cuál es el valor-p?
d. ¿A qué conclusión llega?

15. Las lesiones entre los jugadores de la Liga Mayor de béisbol han aumentado en los últimos años. La expansión de la Liga, de 1992 a 2001, hizo que la lista de nombres aumentara 15%. Sin embargo, la cantidad de jugadores en la lista de inhabilitados por causa de una lesión aumentó 32% en ese mismo periodo (*USA Today*, 8 de julio de 2002). La cuestión a investigar es si los jugadores permanecen en la lista de inhabilitados más tiempo que quienes permanecían en la lista una década antes.

a. Con la media poblacional de la cantidad de días que permanece un jugador en la lista de inhabilitados, formule las hipótesis nula y alternativa que se pueden usar para probar la cuestión a investigar.

b. Tome como datos los siguientes:

	Temporada 2001	Temporada 1992
Tamaño de la muestra	$n_1 = 45$	$n_2 = 38$
Media muestral	$\bar{x}_1 = 60$ días	$\bar{x}_2 = 51$ días
Desviación estándar muestral	$s_1 = 18$ días	$s_2 = 15$ días



¿Cuál es la estimación puntual de la diferencia entre las medias poblacionales de la cantidad de días en la lista de inhabilitados en 2001 y en 1992? ¿Cuál es el porcentaje de incremento en el número de días en la lista de inhabilitados?

c. Use $\alpha = 0.01$. ¿Cuál es la conclusión acerca de la cantidad de días en la lista de inhabilitados? ¿Cuál es el valor-p?

d. ¿Estos datos indican que la Liga Mayor de Béisbol deberá preocuparse por la situación?



SATVerbal

16. El consejo universitario compara las puntuaciones obtenidas en la prueba de aptitudes escolares (SAT, por sus siglas en inglés) de acuerdo con el nivel de enseñanza de los padres de los estudiantes que presentan este examen. La hipótesis de investigación es que los estudiantes cuyos padres tienen un nivel más alto de estudios obtendrán mejores puntuaciones en el SAT. En el 2003 la media general en la prueba oral fue 507 (*The World Almanac 2004*). A continuación se presentan las puntuaciones obtenidas en el examen verbal en dos muestras independientes de estudiantes. La primera muestra corresponde a las puntuaciones de estudiantes cuyos padres tienen una licenciatura. La segunda corresponde a las puntuaciones de estudiantes cuyos padres terminaron la preparatoria pero no tienen una licenciatura.

Padres de los estudiantes			
Con licenciatura		**Con preparatoria**	
485	487	442	492
534	533	580	478
650	526	479	425
554	410	486	485
550	515	528	390
572	578	524	535
497	448		
592	469		

a. Formule las hipótesis pertinentes para determinar si los datos muestrales confirman la hipótesis de que los estudiantes cuyos padres tienen un nivel de enseñanza más alto obtienen mejores puntuaciones en el SAT.

b. Dé la estimación puntual de la diferencia entre las medias de las dos poblaciones.

c. Calcule el valor-p en esta prueba de hipótesis.

d. Con $\alpha = 0.05$, ¿cuál es la conclusión?

17. Merrill Lynch solicita periódicamente a sus clientes evaluaciones sobre la asesoría financiera y los servicios que les presta (2000 Merrill Lynch Client Satisfation Survey). Puntuaciones más altas indican mejor servicio, 7 es la puntuación más alta. A continuación se presentan en forma resumida las puntuaciones dadas a dos consultores financieros por los miembros de dos muestras aleatorias independientes. El consultor A tiene 10 años de experiencia, mientras que el consultor B tiene 1 año de experiencia. Use $\alpha = 0.05$ y realice una prueba para determinar si el consultor que tiene más años de experiencia obtuvo una puntuación más alta.

Consultor A	Consultor B
$n_1 = 16$	$n_2 = 10$
$\bar{x}_1 = 6.82$	$\bar{x}_2 = 6.25$
$s_1 = 0.64$	$s_2 = 0.75$

a. Establezca las hipótesis nula y alternativa.
b. Calcule el valor del estadístico de prueba.
c. ¿Cuál es el valor-*p*?
d. ¿A qué conclusión llega?



SAT

18. Las empresas que se dedican a dar tutoría a estudiantes organizan asesorías, clases y preparación para exámenes con el fin de ayudar a los estudiantes a obtener mejores resultados, como en el examen de aptitudes escolares (SAT, por sus siglas en inglés). Estas empresas aseguran que sus cursos ayudan a los estudiantes a mejorar sus puntuaciones en estos exámenes hasta en un promedio de 120 puntos (*The Wall Street Journal*, 23 de enero de 2003). Un investigador duda de esta aseveración y cree que 120 puntos es una exageración de las empresas para motivar a los estudiantes a tomar los cursos de preparación. En un estudio para evaluar un curso de preparación para dicho examen, los investigadores recogieron datos de las puntuaciones de 35 estudiantes que tomaron un curso y de 48 estudiantes que no tomaron el curso. El archivo SAT del disco compacto contiene los datos de tal estudio.
   a. Formule las hipótesis para probar la suposición de los investigadores de que la mejora en la puntuación del SAT debe ser menor que 120 puntos.
   b. Use $\alpha = 0.05$. ¿Cuál es la conclusión?
   c. Cuál es la estimación puntual de la mejora en la puntuación promedio del SAT obtenida con los cursos. Dé un intervalo de confianza de 95% para la estimación de la mejora.
   d. ¿Qué consejo daría al investigador después de ver el intervalo de confianza?



Inferencias acerca de la diferencia entre dos medias poblacionales: muestras pareadas

Suponga que los empleados de una fábrica usan dos métodos distintos para realizar una determinada tarea. Con objeto de maximizar la producción, la empresa desea identificar el método con el que la media poblacional del tiempo necesario para realizar esta tarea sea menor. Sea μ_1 la media poblacional del tiempo empleando el método 1 y μ_2 la media poblacional del tiempo requerido para realizar la tarea con el método 2. Puesto que no hay ninguna indicación de cuál sea el mejor método, se empieza por suponer que con los dos métodos se obtiene la misma media poblacional del tiempo requerido para realizar la tarea. De esta manera, la hipótesis nula es H_0: $\mu_1 - \mu_2 = 0$. Si se rechaza esta hipótesis se podrá concluir que las medias poblacionales de los tiempos requeridos para realizar la tarea son diferentes con los dos métodos. En tal caso se recomendará el método que proporcione el menor tiempo para la realización de la tarea. Las hipótesis nula y alternativa se expresan como sigue.

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

En la elección del método de muestreo para obtener los datos y probar las hipótesis, se consideran dos diseños alternativos. Uno se basa en muestras independientes y el otro en **muestras por pares**.

1. *Diseño de muestras independientes:* se toma una muestra aleatoria simple de trabajadores y cada uno de ellos usa el método 1. Se toma otra muestra aleatoria simple de trabajado-

res y cada uno de ellos usa el método 2. El procedimiento que se usa para probar la diferencia entre las dos medias es el procedimiento presentado en la sección 10.2.

2. *Diseño de muestras pareadas:* se toma una muestra aleatoria simple de trabajadores. Cada trabajador primero usa uno de los métodos y después usa el otro método. A cada trabajador se le asigna en forma aleatoria el orden en que usará los dos métodos, algunos trabajadores primero usarán el método 1 y otros el método 2.

En el diseño de muestras pareadas los dos métodos se prueban bajo condiciones similares (es decir, con los mismos trabajadores); por tanto, este diseño suele conducir a errores muestrales más pequeños que el diseño de muestras independientes. La razón principal es que en el diseño de muestras pareadas se elimina la variación entre los trabajadores, ya que los mismos trabajadores prueban los dos métodos.

A continuación, con el empleo del diseño de muestras pareadas se demostrará la diferencia entre las medias de los dos métodos de producción. Se emplea una muestra aleatoria de seis trabajadores. En la tabla 10.2 se muestran los tiempos que requirieron los trabajadores para realizar la tarea. Observe que de cada trabajador se obtuvieron dos datos, uno con cada método de producción, también que en la última columna se da, para cada trabajador de la muestra, la diferencia d_i entre los tiempos para realizar la tarea.

Lo principal en el análisis de muestras pareadas es darse cuenta de que únicamente hay que considerar la columna de las diferencias. De manera que se tienen seis datos (0.6, −0.2, 0.5, 0.3, 0.0 y 0.6) que se usarán para analizar la diferencia entre las medias poblacionales de los dos métodos de producción.

Sea μ_d la media de las *diferencias* en la población de trabajadores. Con esta notación, las hipótesis nula y alternativa se expresan como sigue:

$$H_0: \mu_d = 0$$
$$H_a: \mu_d \neq 0$$

Si se rechaza H_0, se concluye que difieren las medias poblacionales de los tiempos requeridos para realizar la tarea con los dos métodos.

Fuera del uso de la notación d, las fórmulas para la media muestral y para la desviación estándar muestral son las mismas que se han usado ya antes en este libro.

La notación d sirve para recordar que las muestras pareadas proporcionan datos que son *diferencias*. A continuación se calcula la media y la desviación estándar de las seis diferencias que se presentan en la tabla 10.2

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{1.8}{6} = 0.30$$

$$s_d = \sqrt{\frac{\Sigma(d_i - \bar{d})^2}{n - 1}} = \sqrt{\frac{0.56}{5}} = 0.335$$

TABLA 10.2 TIEMPOS PARA LA REALIZACIÓN DE UNA TAREA. DISEÑO DE MUESTRAS PAREADAS



Matched

Trabajador	Tiempo para realizar la tarea con el método 1 (minutos)	Tiempo para realizar la tarea con el método 2 (minutos)	Diferencia entre los tiempos (d_i)
1	6.0	5.4	0.6
2	5.0	5.2	−0.2
3	7.0	6.5	0.5
4	6.2	5.9	0.3
5	6.0	6.0	0.0
6	6.4	5.8	0.6

Si la muestra es grande no es necesaria la suposición de que la población tiene una distribución normal. En los capítulos 8 y 9 se presentaron los lineamientos sobre el tamaño de la muestra para usar la distribución t.

Como la muestra es pequeña, $n = 6$, es necesario suponer que la población de las diferencias tiene una distribución normal. Esta suposición es necesaria para usar la distribución t en la prueba de hipótesis y para calcular una estimación por intervalo. Con esta suposición, el estadístico de prueba siguiente tiene una distribución t con $n - 1$ grados de libertad.

ESTADÍSTICO DE PRUEBA PARA PRUEBAS DE HIPÓTESIS CON MUESTRAS PAREADAS

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \tag{10.9}$$

Una vez que la diferencia de datos es calculada, el procedimiento para la distribución t para las muestras por pares, es el mismo que para una estimación de una población y la prueba de hipótesis descritas en los capítulos 8 y 9.

A continuación se usará la ecuación (10.9) para probar las hipótesis H_0: $\mu_d = 0$ y H_a: $\mu_d \neq 0$, usando $\alpha = 0.05$. El estadístico de prueba se calcula sustituyendo en la ecuación (10.9) los resultados muestrales, $\bar{d} = 0.30$, $s_d = 0.335$ y $n = 6$.

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{0.30 - 0}{0.335/\sqrt{6}} = 2.20$$

Ahora se calculará el valor-p para esta prueba de dos colas. Como $t = 2.20 > 0$, el estadístico de prueba se encuentra en la cola superior de la distribución t. Como $t = 2.20$, el área en la cola superior a la derecha del estadístico de prueba se encuentra usando la tabla de la distribución t con $= n - 1 = 6 - 1 = 5$ grados de libertad. A continuación se copia la información correspondiente al renglón de la distribución t para 5 grados de libertad.

Área en la cola superior	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
Valor *t* (5 gl)	0.920	1.476	2.015	2.571	3.365	4.032

$t = 2.20$

Como se ve, el área en la cola superior está entre 0.05 y 0.025. Por tratarse de una prueba de dos colas, se duplica este valor y se concluye que el valor-p está entre 0.10 y 0.05. Este valor-p es mayor a $\alpha = 0.05$. Por ende, no se rechaza la hipótesis nula H_0: $\mu_d = 0$. Con Excel o Minitab y los datos de la tabla 10.2 se halla el valor exacto del valor-p, que es 0.080.

Además, también se obtiene un intervalo de confianza para estimar la diferencia entre las dos medias poblacionales usando la metodología presentada en el capítulo 8 para una sola población. A continuación, el cálculo para obtener un intervalo de confianza de 95%.

$$\bar{d} \pm t_{0.025}\frac{s_d}{\sqrt{n}}$$

$$0.3 \pm 2.571\left(\frac{0.335}{\sqrt{6}}\right)$$

$$0.3 \pm 0.35$$

Por tanto, el margen de error es 0.35 y el intervalo de 95% de confianza para estimar la diferencia entre las medias poblacionales de los dos métodos de producción es el intervalo que va de −0.05 minutos a 0.65 minutos.

NOTAS Y COMENTARIOS

1. En el ejemplo presentado en esta sección, los trabajadores realizan la tarea primero con un método y luego con el otro. Este ejemplo ilustra un diseño de muestras pareadas en el que de cada elemento de la muestra se obtienen dos datos. Para obtener el par de datos, también se emplean elementos diferentes pero "similares". Por ejemplo, un trabajador en una ubicación forma pareja con otro en diferente ubicación (con similitud en edad, género, experiencia, nivel de estudio, etc.). De las parejas de trabajadores se obtendrán los datos de las diferencias a ser usados en el análisis de muestras pareadas.
2. Con el método de muestras pareadas para obtener inferencias sobre dos medias poblacionales, por lo general, se obtienen mejores resultados que con el método de muestras independientes. Sin embargo, en muchas aplicaciones no se logran formar pares o el tiempo y el costo requeridos son excesivos. En tales casos deberá usarse el método de muestras independientes.

Ejercicios

Métodos



19. Considere la prueba de hipótesis siguiente

$$H_0: \mu_d \leq 0$$
$$H_a: \mu_d > 0$$

Los datos siguientes provienen de muestras pareadas tomadas de dos poblaciones.

	Población	
Elemento	**1**	**2**
1	21	20
2	28	26
3	18	18
4	20	20
5	26	24

a. Calcule la diferencia en cada elemento.
b. Calcule $\bar{d}$
c. Calcule la desviación estándar s_d.
d. Realice una prueba de hipótesis usando $\alpha = 0.05$. ¿Cuál es la conclusión?

20. Los datos siguientes provienen de muestras pareadas tomadas de dos poblaciones.

	Población	
Elemento	**1**	**2**
1	11	8
2	7	8
3	9	6
4	12	7
5	13	10
6	15	15
7	15	14

a. Estime la diferencia en cada elemento.
b. Calcule $\bar{d}$.
c. Calcule la desviación estándar s_d.
d. Dé la estimación puntual de la diferencia entre las dos medias poblacionales.
e. Dé un intervalo de 95% de confianza para estimar la diferencia entre las dos medias poblacionales.

Aplicaciones



21. Una empresa de investigación de mercado emplea una muestra de individuos para calificar el potencial de compra de un determinado producto antes y después de que los individuos vean un comercial de televisión acerca del mismo. La calificación del potencial de compra se hace con una escala del 0 al 10, con los valores más altos indicando un mayor potencial de compra. En la hipótesis nula se establece que la media de las calificaciones de "después" será menor o igual a la media de las calificaciones "antes". El rechazo de esta hipótesis indica que el comercial mejora la media de la calificación al potencial de compra. Use $\alpha = 0.05$ y los datos de la tabla siguiente para probar esta hipótesis y haga un comentario sobre la utilidad del comercial.

	Calificación al potencial de compra			Calificación al potencial de compra	
Individuos	**Después**	**Antes**	**Individuos**	**Después**	**Antes**
1	6	5	5	3	5
2	6	4	6	9	8
3	7	7	7	7	5
4	4	3	8	6	6



22. Datos sobre las ganancias por acción en los que se comparan las ganancias en un trimestre con las ganancias del trimestre anterior se encuentran en el archivo titulado Earning 2005 del disco compacto. Dé un intervalo de confianza de 95% para estimar la diferencia entre las medias poblacionales del trimestre dado frente a las del trimestre anterior. ¿Las ganancias aumentaron?

23. En un estudio del Bank of America sobre el gasto de los consumidores, se recogieron datos sobre las cantidades pagadas con tarjetas de crédito en seis categorías diferentes: transporte, supermercado, cenar fuera, gastos para el hogar, muebles para el hogar, vestido y diversión (*U.S. Airways Attaché*, diciembre de 2003). Suponga que con datos de 43 tarjetas de crédito se identifican las cantidades anuales que se gastaron en supermercado (población 1) y en cenar fuera (población 2). A partir de las diferencias, la media muestral de éstas fue $\bar{d} = \$850$ y la desviación estándar muestral fue $s_d = \$1\ 123$.
    a. Formule las hipótesis nula y alternativa para probar que no hay diferencia entre la media poblacional de los gastos en supermercado pagados con tarjeta de crédito y la media poblacional de los gastos en cenar fuera pagados con tarjeta de crédito.
    b. Con 0.05 como nivel de significancia, ¿se puede concluir que hay diferencia entre las medias poblacionales? ¿Cuál es el valor-*p*?
    c. ¿En qué categoría, supermercado o cenar fuera, es mayor la media poblacional de los gastos anuales pagados con tarjeta de crédito? Dé la estimación puntual de una diferencia entre las medias poblacionales. Proporcione un intervalo de confianza de 95% para estimar la diferencia entre estas medias poblacionales.



24. Las personas que viajan por avión, suelen elegir de qué aeropuerto salir con base en el costo del vuelo. Para determinar de qué aeropuerto es más costoso salir, si de Dayton, Ohio; o de Louisville, Kentucky, se recolectan datos (en dólares) de una muestra de vuelos a ocho ciudades partiendo de estos dos aeropuertos (*The Cincinnati Enquirer*, 19 de febrero de 2006). Un investigador sostiene que es mucho más costoso partir de Dayton, que de Louisville. Use los datos muestrales para ver si favorecen tal afirmación. Como nivel de significancia use $\alpha = 0.05$.

Destino	Dayton	Louisville
Chicago-O'Hare	$319	$142
Grand Rapids, Michigan	192	213
Portland, Oregon	503	317
Atlanta	256	387
Seattle	339	317
South Bend, Indiana	379	167
Miami	268	273
Dallas–Ft. Worth	288	274

25. En los últimos tiempos hay una cantidad cada vez mayor de opciones de entretenimiento que compiten por el tiempo de los consumidores. En 2004, la televisión por cable y el radio superaron a la televisión abierta, a la música grabada y a los periódicos, convirtiéndose en los medios de entretenimiento más usados. Con una muestra de 15 individuos se obtienen los datos de las horas por semana que ven televisión por cable y de las horas por semana que escuchan la radio.



TVRadio

Individuos	Televisión	Radio	Individuos	Televisión	Radio
1	22	25	9	21	21
2	8	10	10	23	23
3	25	29	11	14	15
4	22	19	12	14	18
5	12	13	13	14	17
6	26	28	14	16	15
7	22	23	15	24	23
8	19	21			

a. Use como nivel de significancia 0.05 y haga una prueba para la diferencia entre las medias poblacionales de la cantidad de horas de televisión por cable y de la cantidad de horas de radio.
b. ¿Cuál es la media muestral de la cantidad de horas por semana empleadas en ver televisión por cable?
c. ¿Cuál es la media muestral de la cantidad de horas por semana empleadas en escuchar radio? ¿Cuál de estos medios tiene mayor uso?

26. StreetInsider.com presenta las ganancias por acción, en 2002, en una muestra de empresas importantes (12 de febrero de 2003). Antes de 2002, analistas financieros pronosticaron las ganancias por acción de estas empresas. Use los datos siguientes para estudiar las diferencias entre las ganancias reales por acción y las estimaciones de los analistas.



Earnings

Empresa	Real	Estimación
AT&T	1.29	0.38
American Express	2.01	2.31
Citigroup	2.59	3.43
Coca-Cola	1.60	1.78
DuPont	1.84	2.18
ExxonMobil	2.72	2.19
General Electric	1.51	1.71
Johnson & Johnson	2.28	2.18
McDonald's	0.77	1.55
Wal-Mart	1.81	1.74

a. Use $\alpha = 0.05$ y pruebe si existe diferencia entre la media poblacional real y la media poblacional estimada de las ganancias por acción. ¿Cuál es el valor-p? ¿A qué conclusión se llega?
b. Dé una estimación puntual de la diferencia entre las dos medias. ¿Tienden los analistas a subestimar o a sobrestimar las ganancias?
c. Con 95% de confianza, ¿cuál es el margen de error en la estimación del inciso b? De acuerdo con esta información, ¿qué recomendaría?

27. Un fabricante produce dos modelos de una lijadora, de lujo y estándar. Los precios de venta de una muestra de distribuidores minoristas se presentan a continuación.

	Precio ($)			Precio ($)	
Minorista	De lujo	Estándar	Minorista	De lujo	Estándar
1	39	27	5	40	30
2	39	28	6	39	34
3	45	35	7	35	29
4	38	30			

a. En los precios sugeridos por el fabricante para los dos modelos, la diferencia es de $10. Use como nivel de significancia 0.05 y pruebe que la diferencia media entre los precios de los dos modelos es $10.
b. Dé un intervalo de 95% de confianza para la diferencia media entre los precios de los dos modelos.



10.4 Inferencias acerca de la diferencia entre dos proporciones poblacionales

Sea p_1 una proporción de la población 1 y p_2 una proporción de la población 2, a continuación se considerarán inferencias acerca de la diferencia entre dos proporciones poblacionales: $p_1 - p_2$. Para las inferencias acerca de estas diferencias, se seleccionan dos muestras aleatorias independientes, una de n_1 unidades de la población 1 y otra de n_2 unidades de la población 2.

Estimación por intervalo para $p_1 - p_2$

En el ejemplo siguiente se mostrará cómo calcular un margen de error y una estimación por intervalo para la diferencia entre dos proporciones poblacionales.

Una empresa que se dedica a elaborar declaraciones de impuestos desea comparar la calidad del trabajo que se realiza en dos de sus oficinas regionales. Con muestras aleatorias de las declaraciones de impuestos elaboradas en dichas oficinas y verificando la exactitud de las declaraciones, la empresa podrá estimar la proporción de declaraciones con errores elaboradas en cada una de estas oficinas. Interesa conocer la diferencia entre las proporciones siguientes:

p_1 = proporción de declaraciones erróneas en la población 1 (oficina 1)
p_2 = proporción de declaraciones erróneas en la población 2 (oficina 2)
$\bar{p}_1$ = proporción muestral en una muestra aleatoria simple de la población 1
$\bar{p}_2$ = proporción muestral en una muestra aleatoria simple de la población 2

La diferencia entre las dos proporciones poblacionales está dada por $p_1 - p_2$. La estimación puntual de $p_1 - p_2$ es la siguiente.

INFERENCIAS ACERCA DE LA DIFERENCIA ENTRE DOS PROPORCIONES POBLACIONALES

$$\bar{p}_1 - \bar{p}_2 \qquad \textbf{(10.10)}$$

Por ende, el estimador puntual de la diferencia entre dos proporciones poblaciones es la diferencia entre las proporciones muestrales de dos muestras aleatorias simples independientes.

Como ocurre con otros estimadores puntuales, $\bar{p}_1 - \bar{p}_2$ tiene una distribución muestral que refleja los valores que podría tomar $\bar{p}_1 - \bar{p}_2$ si se tomaran repetidas muestras aleatorias simples independientes. La media de esta distribución muestral es $p_1 - p_2$ y el error estándar de $\bar{p}_1 - \bar{p}_2$ es el siguiente.

ERROR ESTÁNDAR DE $\bar{p}_1 - \bar{p}_2$

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} \qquad \textbf{(10.11)}$$

Si los tamaños de las muestras son suficientemente grandes para que n_1p_1, $n_1(1 - p_1)$, n_2p_2 y $n_2(1 - p_2)$ sean todos mayores o iguales que 5, la distribución muestral de $\bar{p}_1 - \bar{p}_2$ puede ser aproximada por una distribución normal.

Como ya se indicó antes, una estimación por intervalo está dada por una estimación puntual ± un margen de error. En la estimación de la diferencia entre dos proporciones poblacionales, una estimación por intervalo toma la forma siguiente:

$$\bar{p}_1 - \bar{p}_2 \pm \text{Margen de error}$$

Aproximando la distribución muestral de $\bar{p}_1 - \bar{p}_2$ mediante una distribución normal, se podrá usar como margen de error $z_{\alpha/2}\sigma_{\bar{p}_1 - \bar{p}_2}$, Sin embargo, $\sigma_{\bar{p}_1 - \bar{p}_2}$ como es dada por la ecuación (10.11) no se puede usar directamente porque no se conoce ninguna de las dos proporciones poblacionales p_1 y p_2. Usando la proporción muestral $\bar{p}_1$ para estimar p_1 y la proporción muestral $\bar{p}_2$ para estimar p_2, el margen de error queda como sigue.

$$\text{Margen de error} = z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}} \qquad \textbf{(10.12)}$$

La forma general de una estimación por intervalo para la diferencia entre dos proporciones poblacionales es la siguiente.

ESTIMACIÓN POR INTERVALO PARA LA DIFERENCIA ENTRE DOS PROPORCIONES POBLACIONALES

$$\bar{p}_1 - \bar{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}} \qquad \textbf{(10.13)}$$

donde $1 - \alpha$ es el coeficiente de confianza.

De regreso al ejemplo de elaboración de declaraciones de impuestos, se encuentra que de las muestras independientes aleatorias simples de las dos oficinas se obtienen los datos siguientes.

Oficina 1	Oficina 2
$n_1 = 250$	$n_2 = 300$
Número de declaraciones con errores = 35	Número de declaraciones con errores = 27


TaxPrep

Las proporciones muestrales en cada una de las oficinas son las siguientes.

$$\bar{p}_1 = \frac{35}{250} = 0.14$$

$$\bar{p}_2 = \frac{27}{300} = 0.09$$

La estimación puntual de la diferencia entre las proporciones de declaraciones con errores en las dos poblaciones es $\bar{p}_1 - \bar{p}_2 = 0.14 - 0.09 = 0.05$. Entonces se estima que la oficina 1 comete 0.05 o 5% más errores que la oficina 2.

Ahora se puede usar la expresión (10.13) para calcular el margen de error y la estimación por intervalo para la diferencia entre las dos proporciones poblacionales. Con un intervalo de 90% de confianza con $z_{\alpha/2} = z_{0.05} = 1.645$, se tiene

$$\bar{p}_1 - \bar{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

$$0.14 - 0.09 \pm 1.645\sqrt{\frac{0.14(1-0.14)}{250} + \frac{0.09(1-0.09)}{300}}$$

$$0.05 \pm 0.045$$

El margen de error es 0.045 y el intervalo de 90% de confianza es el intervalo que va de 0.005 a 0.095.

Prueba de hipótesis acerca de $p_1 - p_2$

Ahora se considerarán las pruebas de hipótesis acerca de la diferencia entre las proporciones de dos poblaciones. Se verán pruebas que comprenden el caso en que no hay diferencia entre las dos proporciones poblacionales. En tal caso, las tres formas de las pruebas de hipótesis son las siguientes:

En todas las hipótesis consideradas se usa 0 como la diferencia de interés.

$$H_0\colon p_1 - p_2 \geq 0 \qquad H_0\colon p_1 - p_2 \leq 0 \qquad H_0\colon p_1 - p_2 = 0$$
$$H_a\colon p_1 - p_2 < 0 \qquad H_a\colon p_1 - p_2 > 0 \qquad H_a\colon p_1 - p_2 \neq 0$$

Si se supone que H_0, considerada como igualdad, es verdadera, se tiene $p_1 - p_2 = 0$, que equivale a decir que dichas proporciones poblacionales son iguales, $p_1 = p_2$.

El estadístico de prueba se basará en la distribución muestral del estimador puntual $\bar{p}_1 - \bar{p}_2$ En la ecuación (10.11), se mostró que el error estándar de $\bar{p}_1 - \bar{p}_2$ está dado por

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Si se supone que H_0 es verdadera como igualdad, las proporciones poblacionales son iguales y $p_1 = p_2 = p$. En este caso $\sigma_{\bar{p}_1 - \bar{p}_2}$ se convierte en

ERROR ESTÁNDAR DE $\bar{p}_1 - \bar{p}_2$ CUANDO $p_1 = p_2 = p$

$$\sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \textbf{(10.14)}$$

Como no se conoce p, se combinan los estimadores puntuales de las dos muestras ($\bar{p}_1$ y $\bar{p}_2$) con objeto de obtener un solo estimador puntual de p.

ESTIMADOR COMBINADO DE p CUANDO $p_1 = p_2 = p$

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} \quad \textbf{(10.15)}$$

El **estimador combinado de *p*** es un promedio ponderado de $\bar{p}_1$ y $\bar{p}_2$.

Al sustituir p por $\bar{p}$ en la ecuación (10.14) se obtiene una estimación del error estándar de $\bar{p}_1 - \bar{p}_2$. Dicha estimación del error estándar se usa en el estadístico de prueba. La fórmula general del estadístico de prueba para una prueba de hipótesis acerca de la diferencia entre dos proporciones poblacionales es el estimador puntual dividido entre la estimación de $\sigma_{\bar{p}_1-\bar{p}_2}$.

ESTADÍSTICO DE PRUEBA PARA PRUEBAS DE HIPÓTESIS ACERCA DE $p_1 - p_2$

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \textbf{(10.16)}$$

Tal estadístico de prueba se usa con muestras grandes, en las que n_1p_1, $n_1(1 - p_1)$, n_2p_2 y $n_2(1 - p_2)$, sean todos mayores o iguales que 5.

De nuevo al ejemplo de la empresa que se dedica a elaborar declaraciones de impuestos, suponga que la empresa desea realizar una prueba de hipótesis para determinar si las proporciones de errores en las dos oficinas son diferentes. Para esto, se requiere una prueba de hipótesis de dos colas. Las hipótesis nula y alternativa son las siguientes:

$$H_0\colon p_1 - p_2 = 0$$
$$H_a\colon p_1 - p_2 \neq 0$$

Si se rechaza H_0, la empresa concluiría que la proporción de errores que se cometen en las dos oficinas es distinta. Como nivel de significancia se usará $\alpha = 0.10$.

En los datos muestrales recogidos previamente se encuentra que $\bar{p}_1 = 0.14$ en la muestra de $n_1 = 250$ declaraciones de la oficina 1 y $\bar{p}_2 = 0.09$ en la muestra de $n_2 = 300$ declaraciones en la muestra de la oficina 2. A continuación se calculará la estimación combinada de p.

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} = \frac{250(0.14) + 300(0.09)}{250 + 300} = 0.1127$$

Con la estimación combinada y la diferencia entre las proporciones muestrales, se obtiene el valor del estadístico de prueba como se indica a continuación.

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.14 - 0.09)}{\sqrt{0.1127(1 - 0.1127)\left(\frac{1}{250} + \frac{1}{300}\right)}} = 1.85$$

Para calcular el valor-p para esta prueba de dos colas, se observa, primero, que $z = 1.85$ se encuentra en la cola superior de la distribución normal estándar. A partir de $z = 1.85$ y la tabla de la distribución normal estándar, se encuentra que el área en la cola superior es $1.0000 - 0.9678 = 0.0322$. Multiplicando esta área por dos, dado que se trata de una prueba de dos colas, se encuentra que el valor-p es $2(0.0322) = 0.0644$. Como el valor-p es menor a $\alpha = 0.10$, se rechaza H_0 para el nivel de significancia 0.10. La empresa concluye que las proporciones de errores de las dos oficinas difieren. La conclusión de esta prueba de hipótesis es consistente con los resultados de la estimación por intervalo que se calculó antes y que indicaban que la diferencia entre las proporciones poblacionales en las dos oficinas estaba entre 0.005 y 0.095, siendo la oficina 1 la que tenía una mayor tasa de errores

Ejercicios

Métodos

28. Considere los resultados siguientes obtenidos de muestras independientes tomadas de dos poblaciones.

Muestra 1	Muestra 2
$n_1 = 400$	$n_2 = 300$
$\bar{p}_1 = 0.48$	$\bar{p}_2 = 0.36$

a. ¿Cuál es la estimación puntual de la diferencia entre las dos proporciones poblacionales?
b. Dé un intervalo de 90% de confianza para la diferencia entre las dos proporciones poblacionales.
c. Proporcione un intervalo de 95% de confianza para la diferencia entre las dos proporciones poblacionales.



29. Considere la prueba de hipótesis

$$H_0: p_1 - p_2 \leq 0$$
$$H_a: p_1 - p_2 > 0$$

Los siguientes resultados se obtuvieron de muestras independientes tomadas de las dos poblaciones.

Muestra 1	Muestra 2
$n_1 = 200$	$n_2 = 300$
$\bar{p}_1 = 0.22$	$\bar{p}_2 = 0.16$

a. ¿Cuál es el valor-p?
b. Usando $\alpha = 0.05$, ¿cuál es la conclusión en esta prueba de hipótesis?

Aplicaciones



30. En una encuesta de *BusinessWeek*/Harris se pidió a los ejecutivos de empresas grandes su opinión acerca de sus perspectivas económicas para el futuro. Una de las preguntas era: ¿Piensa usted que en los próximos 12 meses aumentará en su empresa el número de empleados de tiempo completo? En esa encuesta 220 de 400 ejecutivos contestaron sí, mientras que en la encuesta realizada el año anterior, 192 de 400 respondieron sí. Encuentre un intervalo de confianza de 95% para estimar la diferencia entre las proporciones en estas dos encuestas. Dé su interpretación de la estimación por intervalo.

31. En los últimos años ha aumentado el número de personas que emplean Internet para buscar noticias sobre política. Los sitios Web sobre política suelen pedir a los usuarios que den sus opiniones participando en encuestas en línea. Pew Research Center realizó un estudio para conocer la participación de republicanos y demócratas en las encuestas en línea. Se obtuvieron los resultados muestrales siguientes.

Partido político	Tamaño de la muestra	Participantes en encuestas en línea
Republicanos	250	115
Demócratas	350	98

a. Calcule la estimación puntual de la proporción de republicanos que indicaron participar en encuestas en línea. Calcule la estimación puntual de demócratas.
b. ¿Cuál es la estimación puntual de la diferencia entre las dos proporciones poblacionales?
c. Con 95% de confianza, ¿cuál es el margen de error?
d. Representantes de la industria científica de sondeo opinan que la profusión de encuestas en línea puede confundir a las personas. ¿Está usted de acuerdo con esto? Use un intervalo de confianza de 95% para estimar la diferencia entre las proporciones poblacionales de republicanos y demócratas para justificar su respuesta.

32. En un estudio de la American Automobile Association se estudió si era más probable que conductores hombres o mujeres se detuvieran para solicitar indicaciones sobre cómo llegar a una dirección (AAA, enero de 2006). En el estudio se preguntaba: “Si usted y su cónyuge van en su automóvil y se pierden, ¿se detiene para preguntar por la dirección que busca?” En una muestra representativa se encontró que 300 de 811 mujeres dijeron que sí se detenían para preguntar y 255 de 750 hombres dijeron que sí se detenían para preguntar.
a. La hipótesis de investigación afirmaba que era más probable que las mujeres se detuvieran para preguntar por la dirección. Formule las hipótesis nula y alternativa para este estudio.
b. ¿Cuál es el porcentaje de mujeres que dijeron detenerse para preguntar por la dirección?
c. ¿Cuál es el porcentaje de hombres que dijeron detenerse para preguntar por la dirección?
d. Pruebe la hipótesis usando $\alpha = 0.05$. ¿Cuál es el valor-p y cuál es la conclusión a la que esperaría usted que llegara la asociación?

33. Las máquinas tragamonedas son el juego preferido en los casinos de Estados Unidos (*Harrah´s Survey 2002: Profile of the American Gambler*). Los siguientes datos muestrales dan el número de hombres y de mujeres para los que su juego favorito son las máquinas tragamonedas.

	Mujeres	Hombres
Tamaño de la muestra	320	250
Juego favorito: máquinas tragamonedas	256	165

a. Suministre una estimación puntual de la proporción de mujeres que consideran a las máquinas tragamonedas su juego favorito.
b. Dé una estimación puntual de la proporción de hombres que consideran a las máquinas tragamonedas su juego favorito.
c. Dé un intervalo de 95% de confianza para estimar la diferencia entre la proporción de mujeres y la de hombres que consideran a las máquinas tragamonedas su juego favorito.

34. El Bureau of Transportation de Estados Unidos vigila la puntualidad de la llegada de los vuelos de las 10 principales aerolíneas de ese país (*The Wall Street Journal*, 4 de marzo de 2003). Los vuelos que llegan con no más de 15 minutos de retraso se consideran a tiempo. Los siguientes son datos estadísticos del Bureau pertenecientes a enero de 2001 y a enero de 2002.

Enero 2001	En una muestra de 924 vuelos, 742 llegaron a tiempo.
Enero 2002	En una muestra de 842 vuelos, 714 llegaron a tiempo.

a. Dé una estimación puntual de la proporción de vuelos que llegaron a tiempo en 2001.
b. Suministre una estimación puntual de la proporción de vuelos que llegaron a tiempo en 2002.
c. Sea p_1 la proporción poblacional de los vuelos que llegaron a tiempo en 2001 y p_2 la proporción poblacional de los vuelos que llegaron a tiempo en 2002. Plantee las hipótesis a probar para determinar si la puntualidad de las principales líneas aéreas mejoró en este periodo de un año.
d. Si $\alpha = 0.01$, ¿cuál es su conclusión?

35. En una prueba de calidad de dos comerciales de televisión, cada comercial se mostró, en áreas separadas de prueba, seis veces en una semana. A la semana siguiente se realizó una encuesta telefónica para identificar a individuos que habían visto los comerciales. A estas personas se les pidió su opinión sobre cuál era el principal mensaje de estos comerciales. Se obtuvieron los siguientes resultados.

	Comercial A	Comercial B
Número de personas que vio el comercial	150	200
Número de personas que recordaba el mensaje	63	60

a. Use $\alpha = 0.05$ y pruebe la hipótesis de que entre los dos comerciales no hay diferencia en las proporciones poblacionales de personas que recordaron el mensaje.
b. Calcule un intervalo de 95% de confianza para la diferencia entre las proporciones de personas que recordaron el mensaje de las dos poblaciones.

36. Durante el SuperBowl de 2003, un comercial de Miller Lite Beer, conocido como "The Miller Lite Girls", fue uno de los tres más efectivos televisados durante el evento (*USA Today*, 29 de diciembre de 2003). Una encuesta para ver la efectividad de los comerciales, conducida por *USA Today´s* Ad Track, empleó muestras por grupos de edades para ver el efecto de la publicidad en el SuperBowl sobre los distintos grupos de edades. A continuación se presentan los resultados muestrales respecto del comercial de la marca de cerveza.

Edad	Tamaño de la muestra	Le gustó mucho el comercial
Menos de 30 años	100	49
De 30 a 49 años	150	54

a. Formule una prueba de hipótesis para determinar si las proporciones poblacionales de los dos grupos de edades difieren.

b. Dé la estimación puntual de la diferencia entre las dos proporciones poblacionales.
c. Realice la prueba de hipótesis y dé el valor-p. Con $\alpha = 0.05$, ¿cuál es su conclusión?
d. Analice la forma en que el comercial llama la atención del grupo de menor y de mayor edad. ¿Le parecerá a la empresa cervecera que los resultados de esta encuesta le son favorables? Explique.

37. En 2003 *New York Times*/CBS News tomó una muestra de 523 personas adultas que planeaban ir de vacaciones en los próximos seis meses y encontró que 141 pensaban ir en avión (*New York Times News Service*, 2 de marzo de 2003). En un sondeo similar que realizó en mayo de 1993, de los 477 adultos que formaron la muestra, 81 pensaban ir en avión.
    a. Establezca las hipótesis para determinar si en este periodo de 10 años hubo un cambio significativo en la proporción de personas que pensaban ir en avión a sus vacaciones.
    b. Suministre la proporción muestral de las personas que pensaban ir en avión en el estudio de 2003. ¿Y en el de 1993?
    c. Con $\alpha = 0.01$ pruebe si hay diferencia significativa. ¿A qué conclusión llega?
    d. Analice las razones que puedan servir como explicación para esta conclusión.

Resumen

En este capítulo se estudiaron los procedimientos para desarrollar estimaciones por intervalos y para realizar pruebas de hipótesis cuando se tienen dos poblaciones. Primero se mostró cómo hacer inferencias acerca de la diferencia entre dos medias poblacionales con muestras aleatorias simples independientes. Enseguida se estudió el caso en que las desviaciones poblacionales s_1 y s_2 se conocen. Se usó la distribución normal estándar z para obtener una estimación por intervalo que sirvió como estadístico de prueba en la prueba de hipótesis. Después se estudió el caso en que las desviaciones estándar poblacionales no se conocen y se estiman mediante las desviaciones estándar muestrales s_1 y s_2. En este caso se usó la distribución t para obtener una estimación por intervalo que sirvió como estadístico de prueba en la prueba de hipótesis.

A continuación se estudiaron las diferencias entre medias poblacionales con el diseño de muestras pareadas. En el diseño de muestras pareadas, cada elemento proporciona un par de datos, uno de cada población, y la diferencia entre los pares de datos se usa para el análisis estadístico. El diseño de muestras pareadas se suele preferir al diseño de muestras independientes debido a que con el diseño de muestras pareadas se suele mejorar la precisión de la estimación.

Por último, se estudiaron los intervalos de estimación y las pruebas de hipótesis para la diferencia entre dos medias poblacionales.

Glosario

Estimador combinado de *p* Estimador de una proporción poblacional que se obtiene calculando un promedio ponderado de los estimadores obtenidos de dos muestras independientes.

Muestras aleatorias simples independientes Muestras tomadas de dos poblaciones, de manera que los elementos que constituyen una muestra se tomen independientemente de los elementos que constituyen la otra muestra.

Muestras pareadas Muestras en las que un dato de una muestra corresponde a un dato de otra muestra

Fórmulas clave

Estimador puntual de la diferencia entre dos medias poblacionales

$$\bar{x}_1 - \bar{x}_2 \tag{10.1}$$

Error estándar de $\bar{x}_1 - \bar{x}_2$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{10.2}$$

Estimación por intervalo para la diferencia entre dos medias poblacionales: σ_1 y σ_2 conocidas

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{10.4}$$

Estadístico de prueba para pruebas de hipótesis acerca de $\mu_1 - \mu_2$: σ_1 y σ_2 conocidas

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{10.5}$$

Estimación por intervalo para la diferencia entre dos medias poblacionales: σ_1 y σ_2 desconocidas

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{10.6}$$

Grados de libertad: distribución *t* con dos muestras aleatorias independientes

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} \tag{10.7}$$

Estadístico de prueba para pruebas de hipótesis acerca de $\mu_1 - \mu_2$: σ_1 y σ_2 desconocidas

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{10.8}$$

Estadístico de prueba para pruebas de hipótesis con muestras pareadas

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \tag{10.9}$$

Estimación puntual de la diferencia entre dos proporciones poblacionales

$$\bar{p}_1 - \bar{p}_2 \tag{10.10}$$

Error estándar de $\bar{p}_1 - \bar{p}_2$

$$\sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \quad \textbf{(10.11)}$$

Estimación por intervalo de la diferencia entre dos proporciones poblacionales

$$\bar{p}_1 - \bar{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \quad \textbf{(10.13)}$$

Error estándar de $\bar{p}_1 - \bar{p}_2$ cuando $p_1 = p_2 = p$

$$\sigma_{\bar{p}_1-\bar{p}_2} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \textbf{(10.14)}$$

Estimador combinado de p cuando $p_1 = p_2 = p$

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} \quad \textbf{(10.15)}$$

Estadístico de prueba para pruebas de hipótesis acerca de $p_1 - p_2$

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \textbf{(10.16)}$$

Ejercicios complementarios

38. Safegate Foods Inc. está rediseñando las cajas en sus supermercados en todo el país y está probando dos diseños. Estos dos sistemas se instalaron en dos supermercados y se midió el tiempo que tardaban los clientes en pasar por la caja. Los resultados se presentan resumidos en la siguiente tabla.

Sistema A	Sistema B
$n_1 = 120$	$n_2 = 100$
$\bar{x}_1 = 4.1$ minutos	$\bar{x}_2 = 3.4$ minutos
$\sigma_1 = 2.2$ minutos	$\sigma_2 = 1.5$ minutos

Con 0.05 como nivel de significancia realice una prueba de hipótesis para determinar si hay diferencia entre las medias poblacionales del tiempo que tardan los clientes en pasar por la caja con estos dos sistemas. ¿Cuál de los sistemas se preferirá?

39. Las cámaras digitales de tres megapixeles suelen ser ligeras, compactas y fáciles de manejar. Pero, si se desea amplificar o recortar imágenes, quizá se estará dispuesto a gastar un poco más en un modelo que dé mayor resolución. A continuación se presentan datos muestrales de cámaras digitales de tres y cinco megapixeles.



Cinco megapixeles		Tres megapixeles	
Modelos	**Precio $**	**Modelos**	**Precio $**
Nikon 5700	890	Kodak DX4330	280
Olympus C-5050	620	Canon A70	290
Sony DCS-F717	730	Sony DSC P8	370
Olympus C-5050	480	Minolta XI	400
Minolta 7Hi	1060	Sony DSC P72	310
HP 935	450	Nikon 3100	340
Pentax 550	540	Panasonic DMC-LC33	270
Canon S50	500	Pentax S	380
Kyocera TVS	890		
Minolta F300	440		

a. Dé una estimación puntual para la diferencia entre las medias poblacionales de los precios de los dos tipos de cámaras. ¿Qué observa acerca de los precios de las cámaras de cinco megapixeles?
b. Dé un intervalo de 95% de confianza para la diferencia entre las medias poblacionales de los precios de los dos tipos de cámaras.

40. Los fondos mutualistas se clasifican en fondos *con comisión* y *sin comisión*. En los fondos con comisión se requiere que el inversionista pague una cantidad base o un porcentaje de la cantidad invertida en el fondo. En los fondos sin comisión no se requiere este pago inicial. Algunos consejeros financieros aseguran que vale la pena pagar la comisión de los fondos con comisión porque tienen tasas medias de rendimiento mayores que los fondos sin comisión. Se tomaron muestras de 30 fondos mutualistas con comisión y 30 fondos mutualistas sin comisión. Se recogieron los datos sobre el rendimiento anual de estos fondos en un periodo de cinco años. Estos datos se presentan en el conjunto de datos Mutual del disco compacto. Los datos de los cinco primeros fondos con comisión y de los cinco primeros fondos sin comisión se presentan en la tabla siguiente.



Fondos mutualistas con comisión	Rendimiento %	Fondos mutualistas sin comisión	Rendimiento %
American National Growth	15.51	Amana Income Fund	13.24
Arch Small Cap Equity	14.57	Berger One Hundred	12.13
Bartlett Cap Basic	17.73	Columbia International Stock	12.17
Calvert World International	10.31	Dodge & Cox Balanced	16.06
Colonial Fund A	16.23	Evergreen Fund	17.61

a. Formule H_0 y H_a de manera que el rechazo de H_0 lleve a la conclusión de que en este periodo de cinco años los fondos mutualistas con comisión tienen un mayor rendimiento medio anual.
b. Use los 60 fondos mutualistas del conjunto de datos Mutual para realizar la prueba de hipótesis. ¿Cuál es el valor-P? Con $\alpha = 0.05$, ¿cuál es su conclusión?

41. La National Association of Home Builders presenta datos sobre los costos de las remodelaciones más frecuentes que se hacen a casas habitación. A continuación se presentan datos muestrales, dados en miles de dólares, de los dos tipos de remodelación más frecuentes.

Cocina	Recámara	Cocina	Recámara
25.2	18.0	23.0	17.8
17.4	22.9	19.7	24.6
22.8	26.4	16.9	21.0
21.9	24.8	21.8	
19.7	26.9	23.6	

a. Dé una estimación puntual de la diferencia entre las medias poblacionales de los costos de los dos tipos de remodelación.
b. Proporcione un intervalo de 90% de confianza para la diferencia entre estas dos medias poblacionales.

42. Una muestra tomada en 15 zonas metropolitanas del estado de Florida indica los precios que suelen tener las casas habitación en dicho estado (en miles de dólares).



Florida

Zona metropolitana	Enero de 2003	Enero de 2002
Daytona Beach	117	96
Fort Lauderdale	207	169
Fort Myers	143	129
Fort Walton Beach	139	134
Gainesville	131	119
Jacksonville	128	119
Lakeland	91	85
Miami	193	165
Naples	263	233
Ocala	86	90
Orlando	134	121
Pensacola	111	105
Sarasota-Bradenton	168	141
Tallahassee	140	130
Tampa-St. Petersburg	139	129

a. Use muestras pareadas para obtener una estimación puntual del incremento de la media poblacional de los precios, en este periodo de un año, de las casas habitación.
b. Dé un intervalo de 90% de confianza para estimar el incremento medio anual, en este periodo, de los precios de las casas habitación en el estado de Florida.
c. ¿De cuánto fue el incremento porcentual en este periodo de un año?

43. Jupiter Media realizó una encuesta para determinar en qué emplean su tiempo libre las personas (*The Wall Street Journal*, 26 de enero de 2004). Ver la televisión es la actividad más popular para pasar el tiempo libre, tanto de los hombres como de las mujeres. La proporción de hombres y la de mujeres que ven la televisión para descansar en su tiempo libre se estima a partir de los siguientes datos muestrales.

Género	Tamaño de la muestra	Ven televisión
Hombres	800	248
Mujeres	600	156

a. Establezca las hipótesis para probar la diferencia entre la proporción poblacional de hombres y la de mujeres que pasan su tiempo libre viendo televisión.
b. ¿Cuál es la proporción muestral de hombres que pasan su tiempo libre viendo televisión? ¿Cuál es la proporción muestral de mujeres?
c. Lleve a cabo la prueba de hipótesis y calcule el valor-*p*. ¿Cuál es la conclusión con 0.05 como nivel de significancia?
d. Dé el margen de error y un intervalo de 95% de confianza para estimar la diferencia entre las dos medias poblacionales.

44. Una empresa grande de seguros de automóviles toma muestras de personas del sexo masculino, asegurados por la empresa, casados y solteros y determina cuántos hicieron uso del seguro en los tres años anteriores.

Asegurados solteros	Asegurados casados
$n_1 = 400$	$n_2 = 900$
Cantidad que hizo uso del seguro = 76	Cantidad que hizo uso del seguro = 90

a. Use $\alpha = 0.05$. Haga una prueba para determinar si la razón de reclamaciones es diferente entre asegurados solteros y casados.
b. Dé un intervalo de 95% de confianza para la diferencia entre las proporciones de las dos poblaciones.

45. Se realizaron pruebas médicas para probar la resistencia a medicamentos contra la tuberculosis. En Nueva Jersey, de 142 casos, 9 fueron resistentes a los medicamentos. En Texas, de 268 casos, 5 fueron resistentes a los medicamentos. ¿Estos datos indican que existe una diferencia estadísticamente significativa entre la proporción de casos resistentes a los medicamentos en estos dos estados? Use 0.02 como nivel de significancia. ¿Cuál es el valor-*p* y cuál es la conclusión a la que se llega?
46. En julio de 2001 Harris Ad Track Research Service realizó una encuesta para evaluar la efectividad de una importante campaña de publicidad para las cámaras fotográficas Kodak (*USA Today*, 27 de agosto de 2007). En una muestra, de los 430 encuestados, 163 dijeron que la publicidad había sido muy eficiente. En otra muestra, de los 285 entrevistados respecto a otra campaña, 66 opinaron que la publicidad había sido muy eficiente.
   a. Estime la proporción de entrevistados que consideraron muy eficiente la publicidad de las cámaras fotográficas Kodak y la proporción de encuestados que opinó que la otra campaña de publicidad había sido muy eficiente.
   b. Dé un intervalo de 95% de confianza para la diferencia entre estas proporciones.
   c. De acuerdo con los resultados del inciso b, ¿cree que la campaña de publicidad de Kodak haya sido más eficiente que la mayoría de las campañas de publicidad?
47. En junio de 2001, 38% de los administradores de fondos encuestados creían que la tasa de inflación subyacente sería más alta en un año. Un mes después, en un estudio similar, 22% de los administradores de fondos pensaban que la inflación subyacente sería más elevada en un año (*Global Research Highlights*, Merrill Lynch, 20 de julio de 2001). Suponga que tanto en junio como en julio el tamaño de la muestra fue 200 encuestados.
   a. Dé una estimación puntual para la diferencia entre proporciones, en junio y julio, de administradores de fondos que esperaban que la inflación subyacente fuera mayor en un año.
   b. Presente hipótesis de manera que el rechazo de la hipótesis nula permita concluir que las expectativas de inflación disminuyeron de junio a julio.
   c. Pruebe las hipótesis del inciso b con $\alpha = 0.05$. ¿Cuál es la conclusión?

Caso problema Par, Inc.

Par, Inc., es un importante fabricante de equipo de golf. El gerente de Par piensa que la participación de la empresa en el mercado aumentaría con la introducción de una pelota de golf resistente al corte y de alta duración. Para esto el grupo de investigación de Par ha estado probando un recubrimiento que dé a las pelotas resistencia al corte y alta durabilidad. Las pruebas realizadas con el recubrimiento han sido prometedoras.

Uno de los investigadores expresó su preocupación por el efecto del nuevo recubrimiento en la distancia de vuelo de la pelota. Par desea que la nueva pelota, resistente al corte, tenga una distancia de recorrido/distancia de vuelo comparable al de las pelotas de golf actuales. Para comparar la distancia de vuelo de los dos tipos de pelotas, se sometieron 40 pelotas de cada modelo a pruebas de distancia. Las pruebas se realizaron con una máquina lanzadora de pelotas con objeto de que la diferencia entre las distancias de vuelo entre los dos modelos de pelota pudiera atri-

buirse a sus diferencias. Los resultados de las pruebas, dados a la yarda más cercana, se presentan en la tabla. Estos datos se encuentran en el disco compacto que se distribuye con el libro.



Modelo		Modelo		Modelo		Modelo	
Actual	**Nuevo**	**Actual**	**Nuevo**	**Actual**	**Nuevo**	**Actual**	**Nuevo**
264	277	270	272	263	274	281	283
261	269	287	259	264	266	274	250
267	263	289	264	284	262	273	253
272	266	280	280	263	271	263	260
258	262	272	274	260	260	275	270
283	251	275	281	283	281	267	263
258	262	265	276	255	250	279	261
266	289	260	269	272	263	274	255
259	286	278	268	266	278	276	263
270	264	275	262	268	264	262	279

Informe administrativo

1. Formule una prueba de hipótesis que le sirva a Par para comparar las distancias de vuelo de la pelota actual y de la nueva pelota.
2. Analice los datos para dar la conclusión de la prueba de hipótesis. ¿Cuál es el valor-*p* de la prueba? ¿Qué le recomendaría a Par, Inc.?
3. Proporcione un resumen de estadística descriptiva con los datos de cada modelo.
4. Dé un intervalo de 95% de confianza para la media poblacional de cada modelo, y otro similar para la diferencia entre las medias poblacionales de los dos modelos.
5. ¿Ve usted que haya necesidad de tomar muestras más grandes y de hacer más pruebas con las pelotas de golf? Discuta.

Apéndice 10.1 Inferencias acerca de dos poblaciones usando Minitab

Aquí se describe el uso de Minitab para calcular estimaciones por intervalos y para realizar pruebas de hipótesis para la diferencia entre dos medias poblacionales y entre dos proporciones poblacionales. Con Minitab se pueden calcular estimaciones por intervalos y hacer pruebas de hipótesis dentro de un mismo módulo. Es decir, Minitab tiene un mismo procedimiento para los dos tipos de inferencias. En los ejemplos siguientes, se mostrará cómo realizar los cálculos para una estimación por intervalo y para una prueba de hipótesis con las dos mismas muestras. Minitab no cuenta con una rutina para inferencias acerca de la diferencia entre dos medias poblacionales cuando no se conocen las desviaciones estándar poblacionales σ_1 y σ_2.

Diferencia entre dos medias poblacionales: σ_1 y σ_2 conocidas



Se emplearán los datos del ejemplo presentado en la sección 10.2, sobre los saldos en las cuentas de cheques. Los datos de los saldos en las cuentas de cheques en la sucursal Cherry Grove se encuentran en la columna 1 y los correspondientes a la sucursal Beechmont se encuentran en la columna C2. En este ejemplo se usará el procedimiento de Minitab 2-Sample *t* para obtener un intervalo de 95% de confianza para estimar la diferencia entre las medias poblacionales de los saldos en las cuentas de cheques de las dos sucursales. En el resultado de este procedimiento Mi-

nitab da también el valor-p para la prueba de hipótesis: H_0: $\mu_1 - \mu_2 = 0$ frente a H_a: $\mu_1 - \mu_2 \neq 0$. Los pasos necesarios para realizar este procedimiento se indican a continuación.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Basic Statistics**
Paso 3. Elegir **2-Sample t**
Paso 4. Cuando aparezca el cuadro de diálogo 2-Sample *t* (Test and Confidence Interval)
Seleccionar **Samples in different columns**
Ingresar C1 en el cuadro **First**
Ingresar C2 en el cuadro **Second**
Seleccionar **Options**
Paso 5. Cuando aparezca el cuadro de diálogo 2-Sample *t*-Options:
Ingresar 95 en el cuadro **Confidence level**
Ingresar 0 en el cuadro **Test difference**
Ingresar *not equal* en el cuadro **Alternative**
Clic en **OK**
Paso 6. Clic en **OK**

La estimación por intervalo de confianza de 95% es el intervalo que va de $37 a $193, como se vio en la sección 10.2. El valor-p = 0.005 indica que la hipótesis nula de que las medias poblacionales son iguales puede rechazarse para el nivel de significancia 0.01. El paso 5 puede modificarse para otras aplicaciones con diferentes niveles de confianza, distintos valores hipotéticos y diversas formas de la prueba de hipótesis.

Diferencia entre dos medias poblacionales con muestras pareadas



Matched

Para ilustrar el procedimiento de muestras pareadas se usarán los datos de la tabla 10.2, sobre los tiempos necesarios para realizar una tarea. Los tiempos con el método 1 se ingresan en la columna C1 y los tiempos con el método 2 se ingresan en la columna C2. Los pasos a seguir al usar Minitab para una prueba de muestras por pares son los siguientes:

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Basic Statistics**
Paso 3. Elegir **Paired t**
Paso 4. Cuando aparezca el cuadro de diálogo Paired t (Test and Confidence Interval)
Seleccionar **Samples in columns**
Ingresar C1 en el cuadro **First sample**
Ingresar C2 en el cuadro **Second sample**
Seleccionar **Options**
Paso 5. Cuando aparezca el cuadro de diálogo Paired t-Options:
Ingresar 95 en el cuadro **Confidence level**
Ingresar 0 en el cuadro **Test mean**
Ingresar *not equal* en el cuadro **Alternative**
Clic en **OK**
Paso 6. Click en **OK**

La estimación por intervalo de 95% de confianza es el intervalo que va de −0.05 a 0.65, como se vio en la sección 10.3. El valor-p = 0.08 indica que la hipótesis nula de que no hay diferencia en los tiempos para realizar la tarea no puede rechazarse para el nivel de significancia α = 0.05. El paso 5 puede modificarse para diferentes niveles de confianza, distintos valores hipotéticos y diversas formas de la prueba de hipótesis.

Diferencia entre dos proporciones poblacionales



TaxPrep

Se emplearán los datos presentados en la sección 10.4 sobre los errores en las declaraciones de impuestos. Los resultados muestrales de 250 declaraciones de impuestos elaboradas en la ofici-

na 1 se encuentran en la columna C1 y los resultados muestrales de 300 declaraciones de impuestos elaboradas en la oficina 2 se encuentran en la columna C2. Sí indica que se encontró un error en la declaración de impuestos y No que no se encontró ningún error. Con el procedimiento que se describe a continuación, se obtiene una estimación por intervalo de 95% de confianza para la diferencia entre las dos proporciones poblacionales, también los resultados de la prueba de hipótesis H_0: $p_1 - p_2 = 0$ y H_a: $p_1 - p_2 \neq 0$.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Basic Statistics**
Paso 3. Elegir **2 Proportions**
Paso 4. Cuando aparezca el cuadro de diálogo *2 Proportions* (*Test and Confidence Interval*):
Seleccionar **Samples in different columns**
Ingresar C1 en el cuadro **First**
Ingresar C2 en el cuadro **Second**
Seleccionar **Options**
Paso 5. Cuando aparezca el cuadro de diálogo *2 Proportions-Options*.
Ingresar 90 en el cuadro **Confidence level**
Ingresar 0 en el cuadro **Test difference**
Ingresar *not equal* en el cuadro **Alternative**
Clic en **OK**
Paso 6. Clic en **OK**

El intervalo de 90% de confianza es el intervalo que va de 0.005 a 0.095, como se vio en la sección 10.4. El valor-p = 0.065 indica que la hipótesis nula de que no hay diferencia entre la tasa de errores se puede rechazar para $\alpha = 0.10$. El paso 5 puede modificarse para diferentes niveles de confianza, distintos valores hipotéticos y diversas formas de la prueba de hipótesis.

En el ejemplo de la elaboración de las declaraciones de impuestos, los datos son cualitativos. Sí y No sirven para indicar si hay o no un error. En el módulo para proporciones, Minitab calcula la proporción de la respuesta que tiene el segundo lugar en las respuestas en orden alfabético. Por tanto, en este ejemplo de la elaboración de las declaraciones de impuestos, Minitab calculará la proporción de respuestas Sí, que es la proporción que se busca.

Si Minitab con el orden alfabético no calcula la proporción de la respuesta de interés, es posible modificar esto. Para ello, seleccione una celda en la columna de los datos, vaya a la barra del menú de Minitab y seleccione *Editor* > *Column* > *Value Order.* Esta secuencia proporcionará la opción de ingresar un orden especificado por el usuario. En la caja *define-an-order* enumere las respuestas colocando en segundo lugar la respuesta de interés. La rutina de Minitab *2 Proportion* facilitará el intervalo de confianza y los resultados de la prueba de hipótesis para la proporción poblacional de interés.

Por último, la rutina de Minitab *2 Proportion* emplea un procedimiento de cálculo distinto al descrito en este libro. Por tanto, quizá los resultados suministrados por Minitab sean ligeramente diferentes, una estimación por intervalo y un valor-p ligeramente diferentes. Sin embargo, los resultados serán muy parecidos y se espera que conduzcan a la misma interpretación y a las mismas conclusiones.

Apéndice 10.2 Inferencias acerca de dos poblaciones usando Excel

Se describirá el uso de Excel para realizar pruebas de hipótesis acerca de la diferencia entre dos medias poblacionales.* Se empieza con las diferencias entre las medias de dos poblaciones cuando se conocen las desviaciones estándar poblacionales σ_1 y σ_2.

*Las herramientas para el análisis de datos de Excel facilitan procedimientos para pruebas de hipótesis acerca de la diferencia entre dos medias poblacionales. Excel no cuenta con una rutina para estimación por intervalos para la diferencia entre dos medias poblacionales ni para inferencias acerca de la diferencia entre dos proporciones poblacionales.

Diferencia entre dos medias poblacionales: σ_1 y σ_2 conocidas



Se empleará el ejemplo de la sección 10.1 sobre las puntuaciones obtenidas en el examen para los dos centros de enseñanza. El rótulo Centro A se encuentra en la celda A1 y el rótulo Centro B se encuentra en la celda B1. Las puntuaciones de los exámenes del centro A se encuentran en las celdas A2:A31 y las puntuaciones de los exámenes del centro B se encuentran en las celdas B2:B41. Se supone que se conocen las desviaciones estándar poblacionales y que son $\sigma_1 = 10$ y $\sigma_2 = 10$. La rutina de Excel solicitará que se ingresen las varianzas, que son $\sigma_1^2 = 100$ y $\sigma_2^2 = 100$. Para realizar una prueba de hipótesis acerca de la diferencia entre dos medias poblacionales se siguen los pasos que se indican a continuación.

Paso 1. Seleccionar el menú **Herramientas**
Paso 2. Elegir **Análisis de datos**
Paso 3. Cuando aparezca el cuadro de diálogo Análisis de datos:
Elegir **Prueba *z* para medias de dos muestras**
Clic en **Aceptar**
Paso 4. Cuando aparezca el cuadro de diálogo Prueba *z* para medias de dos muestras:
Ingresar A1:A31 en el cuadro **Rango para la variable 1**
Ingresar B1:B41 en el cuadro **Rango para la variable 2**
Ingresar 0 en el cuadro **Diferencia hipotética de las medias**
Ingresar 100 en el cuadro **Varianza para la variable 1**
Ingresar 100 en el cuadro **Varianza para la variable 2**
Seleccionar **Rótulos**
Ingresar 0.05 en el cuadro **Alfa**
Seleccionar **Rango de salida** e ingresar C1 en el cuadro correspondiente
Clic en **Aceptar**

El valor-*p* para dos colas se denota $P(Z < = z)$ dos colas y es 0.0978, valor que no permite rechazar la hipótesis nula con $\alpha = 0.05$.

Diferencia entre dos medias poblacionales: σ_1 y σ_2 desconocidas



Se usarán los datos que aparecen en la tabla 10.1 del estudio de la prueba del software. Los datos ya se han ingresado en la hoja de cálculo de Excel, en la que el rótulo Existente se ha ingresado en la celda A1 y el rótulo Nuevo se ha ingresado en la celda B1. Los tiempos requeridos con la tecnología existente se encuentran en las celdas A2:A13 y los tiempos requeridos con el nuevo software se encuentran en las celdas B2:B13. Para realizar una prueba de hipótesis acerca de la diferencia entre dos medias poblacionales cuando no se conocen σ_1 y σ_2 se emplean los siguientes:

Paso 1. Seleccionar el menú **Herramientas**
Paso 2. Elegir **Análisis de datos**
Paso 3. Cuando aparezca el cuadro de diálogo Análisis de datos:
Elegir **Prueba t para dos muestras suponiendo varianzas desiguales**
Clic en **Aceptar**
Paso 4. Cuando aparezca el cuadro de diálogo Prueba t para dos muestras suponiendo varianzas iguales
Ingresar A1:A13 en el cuadro **Rango para la variable 1**
Ingresar B1:B13 en el cuadro **Rango para la variable 2**
Ingresar 0 en el cuadro **Diferencia hipotética entre las medias**
Seleccionar **Rótulos**
Ingresar 0.05 en el cuadro **Alfa**
Seleccionar **Rango de salida** e ingresar C1 en el cuadro correspondiente.
Clic en **Aceptar**

El valor-*p* es el denotado $P(T < =t)$una cola, cuyo valor es 0.017, valor que permite rechazar la hipótesis nula para $\alpha = 0.05$.

Diferencia entre dos medias poblacionales con muestras pareadas



Matched

Para ilustrar este método se usan los pares de datos para la realización de una tarea, que aparecen en la tabla 10.2.

Estos datos se han ingresado en la hoja de cálculo colocando el rótulo Método 1 en la celda A1 y el rótulo Método 2 en la celda B1. Los tiempos requeridos para la realización de la tarea con el método 1 se encuentran en las celdas A2:A7 y los tiempos para la realización de la tarea con el método 2 se encuentran en las celdas B2:B7. En el procedimiento de Excel se emplean los pasos previamente descritos para la prueba-*t* excepto que en el paso 3, cuando aparece el cuadro de diálogo Análisis de datos, hay que elegir **Prueba t para medias de dos muestras por pares**. El rango para la variable 1 es A1:A7 y el rango para la variable 2 es B1:B7. El valor-*p* es el denotado por P(T<=t) dos colas, cuyo valor es 0.08, valor que no permite rechazar la hipótesis nula con $\alpha = 0.05$.



CAPÍTULO 11

Inferencias acerca de varianzas poblacionales

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

LA GENERAL ACCOUNTING OFFICE DE ESTADOS UNIDOS*
WASHINGTON, D. C.

La General Accounting Office de Estados Unidos (GAO) es una organización de auditoría independiente no política perteneciente al área legislativa del gobierno federal. Los auditores de la GAO determinan la efectividad tanto de los programas federales existentes como de los propuestos. Para realizar su labor, los auditores deben ser competentes en revisión de documentos, investigación legislativa y técnicas de análisis estadístico.

En una ocasión los auditores de la GAO estudiaron un programa del Department of Interior que tenía por objeto limpiar los ríos y lagos del país. Como parte de este programa se otorgaron subvenciones a las ciudades pequeñas de Estados Unidos. El Congreso pidió a la GAO que determinara la eficiencia con la que operaba este programa. Con tal objeto, la Gao revisó documentos y visitó varias de las plantas de tratamiento de desechos.

Uno de los objetivos de la auditoría de la GAO era verificar en las plantas que las aguas residuales (desechos tratados) cumplieran determinadas normas. En las auditorías se revisaba, entre otras cosas, datos muestrales sobre contenido de oxígeno, pH y cantidad de sólidos en suspensión en las aguas residuales. Un requisito del programa era que en cada planta diario se realizaran diversas pruebas y que los datos obtenidos se enviaran al departamento de ingeniería del estado. Los datos de la investigación servían para determinar si las características de las aguas residuales se encontraban dentro de límites aceptables.

Así, por ejemplo, examinaron con cuidado los valores promedio de pH. También analizaron la varianza en los valores del pH de las aguas residuales. La prueba de hipótesis acerca de la varianza del pH en la población de aguas residuales fue la siguiente.

$$H_0\colon \sigma^2 = \sigma_0^2$$
$$H_a\colon \sigma^2 \neq \sigma_0^2$$

En esta prueba, σ_0^2 corresponde a la varianza poblacional esperada en los valores de pH de una planta funcionando



Las aguas residuales de esta planta deben estar dentro de un determinado intervalo de valores de pH. © John Boykin/CORBIS.

adecuadamente. Para una de las plantas la hipótesis nula fue rechazada. Análisis más cuidadosos indicaron que en esa planta la varianza del pH era significativamente menor de lo normal.

Los auditores visitaron la planta para revisar el equipo de medición y analizar los resultados con el director de la planta. Encontraron que el equipo no se usaba para la medición de pH debido a que el operador no conocía su funcionamiento. Un ingeniero había informado al operador de los valores de pH aceptables y éste sólo anotaba valores aceptables sin realizar ninguna medición. La varianza inusualmente baja de los datos de esta planta hicieron que se rechazara H_0. La GAO pensó que era probable que otras plantas tuvieran problemas similares y recomendó un programa de capacitación para los operadores, con objeto de mejorar la recolección de datos para el programa de control de la contaminación.

En este capítulo se verá cómo hacer inferencias estadísticas acerca de las varianzas de una o de dos poblaciones. También se presentarán dos nuevas distribuciones, la distribución chi-cuadrada y la distribución *F,* útiles para obtener estimación por intervalos y realizar pruebas de hipótesis para la varianza poblacional.

*Los autores agradecen a William R. Fawle, director de aseguramiento de la calidad de la empresa Colgate-Palmolive por proporcionarles este artículo para *La estadística en la práctica.*

En los cuatro capítulos anteriores se vieron métodos de inferencia estadística para medias y proporciones poblacionales. En este capítulo se extiende dicho estudio a las varianzas poblacionales. Un ejemplo en que la varianza brinda una información importante para tomar una decisión es el caso de un proceso en el que se llenan recipientes con un detergente líquido. La máquina de llenado se ajusta de manera que logre un llenado medio de 16 onzas por envase. Aunque la media de llenado es importante, la varianza en los pesos de llenado también es relevante.

Es decir, aun cuando la máquina de llenado tenga un ajuste adecuado para una media de llenado de 16 onzas, no es de esperar que todos los envases tengan exactamente 16 onzas. Para calcular la varianza muestral de la cantidad de onzas en cada envase se toma una muestra de envases llenos. El valor de la varianza muestral sirve como estimación de la varianza en la población de envases que están siendo llenados en el proceso de producción. Si la varianza muestral es moderada, el proceso continúa. Pero, si la varianza muestral es grande, puede estar ocurriendo por exceso o defecto de llenado, aunque la media sea la correcta. En este caso habrá que reajustar la máquina de llenado con objeto de reducir la varianza de los envases.

En muchos procesos de fabricación controlar la varianza del proceso es de suma importancia para conservar la calidad.

En la primera sección se verán inferencias acerca de la varianza de una sola población. Después, procedimientos para inferencias acerca de varianzas de dos poblaciones.



11.1 Inferencias acerca de una varianza poblacional

La varianza muestral

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} \tag{11.1}$$

es el estimador puntual de la varianza poblacional σ^2. Cuando se hacen inferencias acerca de la varianza poblacional mediante la varianza muestral, la distribución muestral de la cantidad $(n - 1)s^2/\sigma^2$ es de utilidad. Esta distribución muestral se describe como sigue.

DISTRIBUCIÓN MUESTRAL DE $(n - 1)s^2/\sigma^2$

Siempre que de una población normal se tome una muestra aleatoria simple de tamaño n, la distribución muestral de

$$\frac{(n - 1)s^2}{\sigma^2} \tag{11.2}$$

será una distribución chi-cuadrada con $n - 1$ grados de libertad.

La distribución chi-cuadrada parte del muestreo de una población normal.

En la figura 11.1 se muestran algunas de las formas que puede tomar la distribución $(n - 1)s^2/\sigma^2$.

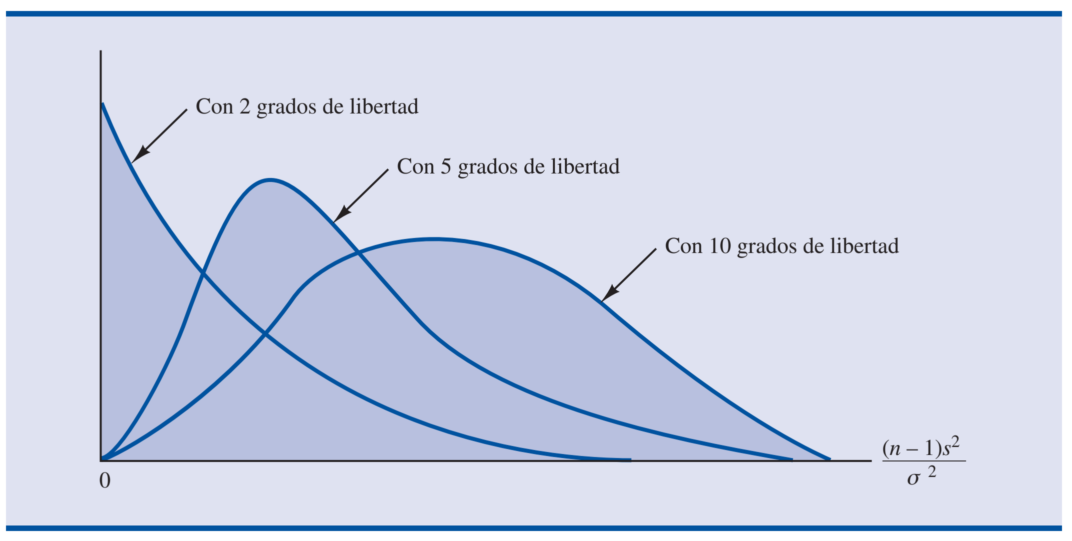
Como ya sabe, siempre que tome una muestra aleatoria simple de tamaño n de una población normal, la distribución muestral de $(n - 1)s^2/\sigma^2$ será una distribución chi-cuadrada, la cual sirve para obtener una estimación por intervalos y realizar pruebas de hipótesis para la varianza poblacional.

Estimación por intervalos

Con el fin de exponer cómo usar la distribución chi-cuadrada para obtener una estimación de la varianza poblacional σ^2 mediante un intervalo de confianza, suponga que desea estimar la varianza poblacional del proceso de llenado citado al comienzo de este capítulo. En una muestra de 20 envases encuentra que la varianza muestral de las cantidades de llenado es $s^2 = 0.0025$. Sin embargo, sabe que no puede esperar que la varianza de 20 envases corresponda al valor exacto de la varianza de toda la población de envases que se llenan en este proceso de producción. Así, deseará obtener una estimación por intervalo para la varianza poblacional.



FIGURA 11.1 EJEMPLOS DE DISTRIBUCIONES MUESTRALES DE $(n - 1)s^2/\sigma^2$ (DISTRIBUCIONES CHI-CUADRADA)



La notación χ^2_α denota el valor de la distribución chi-cuadrada que proporciona un área o probabilidad α a la *derecha* del valor χ^2_α . Por ejemplo, en la figura 11.2, aparece la distribución chi-cuadrada con 19 grados de libertad, en la que $\chi^2_{0.025} = 32.852$ significa que 2.5% de los valores chi-cuadrada se encuentran a la derecha de 32.852 y $\chi^2_{0.975} = 8.907$ significa que 97.5% de los valores chi-cuadrada se encuentran a la derecha de 8.907. Existen tablas que dan las áreas o probabilidades de la distribución chi-cuadrada. Consulte la tabla 11.1 y verifique que los valores de chi-cuadrada con 19 grados de libertad (renglón 19 de la tabla) son correctos. En la tabla 3 del apéndice B se encuentra una tabla más completa con valores chi-cuadrada.

En la gráfica de la figura 11.2 se ve que 0.95 o 95% de los valores chi-cuadrada se encuentran entre $\chi^2_{0.975}$ y $\chi^2_{0.025}$. Es decir, hay un 0.95 de probabilidad de obtener un valor χ^2 tal que.

$$\chi^2_{0.975} \leq \chi^2 \leq \chi^2_{0.025}$$

FIGURA 11.2 UNA DISTRIBUCIÓN CHI-CUADRADA CON 19 GRADOS DE LIBERTAD

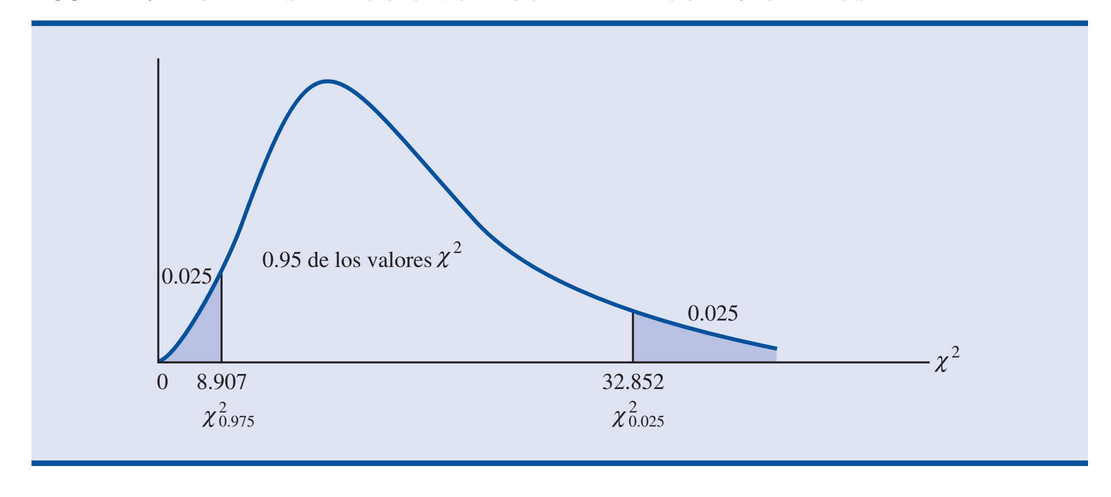
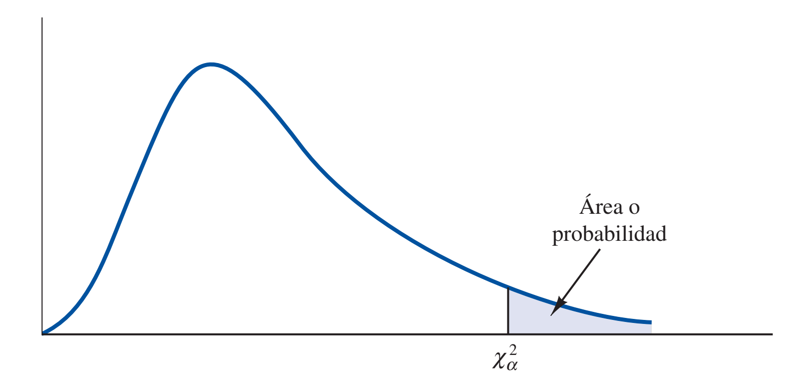


TABLA 11.1 ALGUNOS VALORES DE LA TABLA DE LA DISTRIBUCIÓN CHI-CUADRADA*



Grados de libertad	Área en la cola superior							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

**Nota:* En la tabla 3 del apéndice B se encuentra una tabla más completa.

En la expresión (11.2) se dijo que $(n-1)s^2/\sigma^2$ sigue una distribución chi-cuadrada, por tanto, se puede sustituir χ^2 por $(n-1)s^2/\sigma^2$ y escribir

$$\chi^2_{0.975} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{0.025} \quad \textbf{(11.3)}$$

En efecto, la expresión (11.3) proporciona una estimación por intervalo en la que 0.95 o 95% de todos los valores que puede tomar $(n-1)s^2/\sigma^2$ están en el intervalo que va de $\chi^2_{0.975}$ y $\chi^2_{0.025}$. Ahora es necesario realizar algunas manipulaciones algebraicas a la expresión (11.3) para obtener una estimación por intervalo para la varianza poblacional s^2. Modificando la desigualdad de la izquierda, se tiene

$$\chi^2_{0.975} \leq \frac{(n-1)s^2}{\sigma^2}$$

Por tanto

$$\sigma^2\chi^2_{0.975} \leq (n-1)s^2$$

o

$$\sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{0.975}} \quad \textbf{(11.4)}$$

Con manipulaciones algebraicas similares a la desigualdad de la derecha de la expresión (11.3), se tiene

$$\frac{(n-1)s^2}{\chi^2_{0.025}} \leq \sigma^2 \quad \textbf{(11.5)}$$

Los resultados de las expresiones (11.4) y (11.5) se combinan para obtener

$$\frac{(n-1)s^2}{\chi^2_{0.025}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{0.975}} \quad \textbf{(11.6)}$$

Como la expresión (11.3) es verdadera para 95% de los valores $(n-1)s^2/\sigma^2$, la expresión (11.6) proporciona una estimación por intervalo de confianza de 95% para la varianza poblacional σ^2.

Ahora, de regreso con el problema de dar una estimación por intervalo para la varianza poblacional de las cantidades de llenado, recuerde que en la muestra de 20 envases, la varianza muestral fue $s^2 = 0.0025$. Dado que el tamaño de la muestra es 20, se tienen 19 grados de libertad. Como se muestra en la figura 11.2 ya se determinó que $\chi^2_{0.975} = 8.907$ y $\chi^2_{0.025} = 32.852$. Con estos valores en la expresión (11.6) se obtiene la siguiente estimación por intervalo para la varianza poblacional.

$$\frac{(19)(0.0025)}{32.852} \leq \sigma^2 \leq \frac{(19)(0.0025)}{8.907}$$

Con el fin de obtener un intervalo de confianza para la desviación estándar poblacional, sólo hace falta sacar la raíz cuadrada del límite inferior y del límite superior del intervalo de confianza para la varianza poblacional.

o

$$0.0014 \leq \sigma^2 \leq 0.0053$$

Al sacar la raíz cuadrada de estos valores se obtiene el siguiente intervalo de confianza para la desviación estándar poblacional.

$$0.0380 \leq \sigma \leq 0.0730$$

De esta manera se ha ilustrado el proceso del uso de la distribución chi-cuadrada para obtener una estimación por intervalo para la varianza poblacional y para la desviación estándar poblacional. Observe, que, como se usaron $\chi^2_{0.975}$ y $\chi^2_{0.025}$, el coeficiente de confianza de la estimación por intervalo es 0.95. Extendiendo la expresión (11.6) al caso general, con cualquier coeficiente de confianza, se tiene la siguiente estimación por intervalo para la varianza poblacional.

ESTIMACIÓN POR INTERVALO PARA LA VARIANZA POBLACIONAL

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}} \tag{11.7}$$

donde los valores χ^2 están basados en una distribución chi-cuadrada con $n - 1$ grados de libertad y donde $1 - \alpha$ es el coeficiente de confianza.

Pruebas de hipótesis

Con σ_0^2 para denotar el valor hipotético de la varianza poblacional, las tres formas de una prueba de hipótesis para la varianza poblacional son las siguientes:

$$H_0\colon \sigma^2 \geq \sigma_0^2 \qquad H_0\colon \sigma^2 \leq \sigma_0^2 \qquad H_0\colon \sigma^2 = \sigma_0^2$$
$$H_a\colon \sigma^2 < \sigma_0^2 \qquad H_a\colon \sigma^2 > \sigma_0^2 \qquad H_a\colon \sigma^2 \neq \sigma_0^2$$

Estas tres formas son semejantes a las tres formas, en los capítulos 9 y 10, para las pruebas de hipótesis de una cola y de dos colas para medias poblacionales y proporciones poblacionales.

En una prueba de hipótesis para la varianza poblacional se emplean el valor hipotético de la varianza poblacional σ_0^2 y la varianza muestral s^2 para calcular el valor del estadístico de prueba χ^2. Si la población tiene una distribución normal, el estadístico de prueba es el siguiente:

ESTADÍSTICO DE PRUEBA EN PRUEBAS DE HIPÓTESIS PARA LA VARIANZA POBLACIONAL

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{11.8}$$

donde χ^2 tiene una distribución chi-cuadrada con $n - 1$ grados de libertad.

Una vez calculado el valor del estadístico de prueba χ^2, para determinar si se rechaza la hipótesis nula se emplea el método del valor-p o el método del valor crítico.

Considere ahora el siguiente ejemplo. La St. Louis Metro Bus Company de Estados Unidos, desea dar una imagen de confiabilidad haciendo que sus conductores sean puntuales en los horarios de llegada a las paradas. La empresa desea que haya poca variabilidad en dichos tiempos. En términos de la varianza de los tiempos de llegada a las paradas, la empresa desea que la varianza sea de 4 minutos o menos. Se formula la siguiente prueba de hipótesis para que la empresa pueda determinar si la varianza poblacional en los tiempos de llegada a las paradas es demasiado grande.

$$H_0\colon \sigma^2 \leq 4$$
$$H_a\colon \sigma^2 > 4$$

Suponer, tentativamente, que H_0 sea verdadera, es admitir que la varianza poblacional en los tiempos de llegada a las paradas se encuentra dentro de los lineamientos establecidos por la empresa. La H_0 se rechaza si las evidencias muestrales indican que la varianza poblacional excede los lineamientos de la empresa. En tal caso habrá que tomar medidas para reducir la varianza poblacional. Esta prueba de hipótesis se realiza usando como nivel de significancia $\alpha = 0.05$.



BusTimes

Asuma que en una muestra aleatoria de 24 llegadas a cierta parada en una intersección en el centro de la ciudad, la varianza muestral encontrada es $s^2 = 4.9$. Si la distribución poblacional de los tiempos de llegada a las paradas es aproximadamente normal, el valor del estadístico de prueba es el siguiente:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(24-1)(4.9)}{4} = 28.18$$

En la figura 11.3 se muestra la distribución chi-cuadrada con $n - 1 = 24 - 1 = 23$ grados de libertad. Como ésta es una prueba de la cola superior, el área bajo la curva a la derecha del valor del estadístico de prueba $\chi^2 = 28.18$ es el valor-p de la prueba.

Como ocurre con las tablas de la distribución t, las tablas de la distribución chi-cuadrada no son suficientemente detalladas para permitir determinar con exactitud el valor-p. Sin embargo, con las tablas de la distribución chi-cuadrada se obtiene un intervalo en el que se encuentra el valor-p. Por ejemplo, usando la tabla 11.1 se encuentra la información siguiente para la distribución chi-cuadrada con 23 grados de libertad.

Área en la cola superior	**0.10**	**0.05**	**0.025**	**0.01**
Valor χ^2 (23 *gl*)	32.007	35.172	38.076	41.638

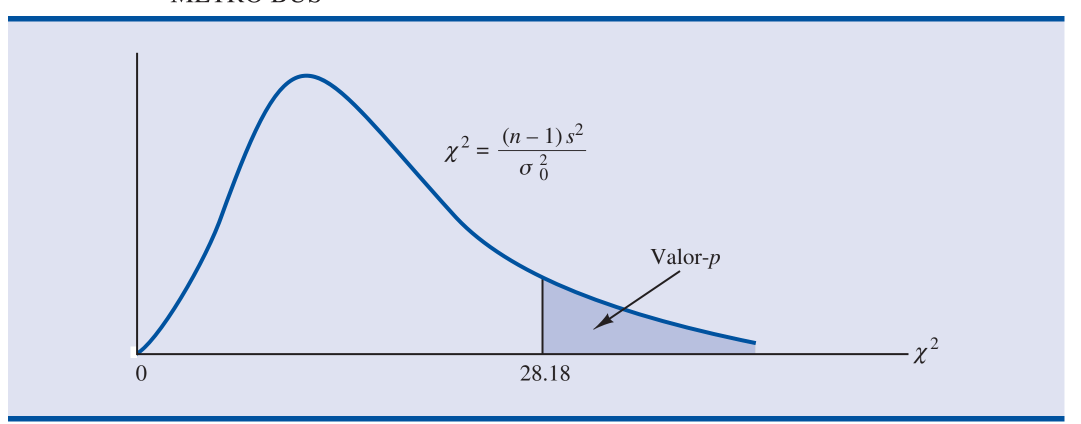
↑
$\chi^2 = 28.18$

Como $\chi^2 = 28.18$ es menor que 32.007, el área en la cola superior (el valor-p) es mayor que 0.10. Como el valor-p es $> \alpha = 0.05$, no se puede rechazar la hipótesis nula. La muestra no lleva a la conclusión de que la varianza en los tiempos de llegada a las paradas sea demasiado grande.

Dada la dificultad para determinar con exactitud el valor-p empleando las tablas de la distribución chi-cuadrada, es útil emplear un paquete de software como Minitab o Excel. En el apéndice F que se encuentra al final del libro se describe cómo calcular los valores-p. En el apéndice se muestra que el exacto valor-p correspondiente a $\chi^2 = 2.18$ es 0.2091.

Como ocurre con los otros procedimientos para pruebas de hipótesis, aquí también es posible emplear el método del valor crítico para obtener la conclusión de la prueba de hipótesis. Así,

FIGURA 11.3 DISTRIBUCIÓN CHI-CUADRADA PARA EL EJEMPLO DE ST. LOUIS METRO BUS



$\alpha = 0.05$, $\chi^2_{0.05}$ proporciona el valor crítico en la cola superior para esta prueba. Empleando la tabla 11.1 y 23 grados de libertad, $\chi^2_{0.05} = 35.172$. De manera que la regla de rechazo para los tiempos de llegada a las paradas es la siguiente:

$$\text{Rechazar } H_0 \text{ si } \chi^2 \geq 35.172$$

Como el valor del estadístico de prueba es $\chi^2 = 28.18$, no se puede rechazar la hipótesis nula.

En la práctica, las pruebas de hipótesis para la varianza poblacional que se encuentran con más frecuencia son pruebas de hipótesis de la cola superior, como la aquí presentada. En situaciones que implican tiempos de llegada, tiempos de producción, pesos de llenado, dimensiones de piezas, etc., varianzas pequeñas son deseables, mientras que varianzas grandes son inaceptables. Al establecer la varianza poblacional máxima permitida, es posible probar la hipótesis nula de que la varianza poblacional es menor o igual que el valor máximo permitido, contra la hipótesis alternativa de que la varianza poblacional es mayor que el valor máximo permitido. Con esta estructura de la prueba, deberán tomarse medidas correctivas, siempre que se rechace la hipótesis nula, lo que indica la presencia de una varianza poblacional demasiado grande.

Como ocurre con la media y la proporción poblacionales, también se realizan pruebas de hipótesis que tienen otras formas. A continuación se presenta una prueba de dos colas para la varianza poblacional considerando una situación que suele presentarse en las oficinas que otorgan licencias para conducir vehículos de motor en Estados Unidos. Históricamente, la varianza en las puntuaciones de los exámenes presentados por las personas que solicitan una licencia para conducir ha sido $\sigma^2 = 100$. Ahora se ha elaborado un nuevo examen con preguntas nuevas. Los administradores de dicha oficina desean que la varianza en las puntuaciones del examen permanezca en los niveles históricos. Para estudiar la varianza en las puntuaciones del nuevo examen se propone la siguiente prueba de hipótesis de dos colas.

$$H_0\colon \sigma^2 = 100$$
$$H_a\colon \sigma^2 \neq 100$$

El rechazo de H_0 indicará que la varianza ha cambiado y que será necesario revisar algunas de las preguntas del nuevo examen para que la varianza en las puntuaciones en este examen sea parecida a la varianza en las puntuaciones del examen anterior. El nuevo examen será aplicado a los integrantes de una muestra de 30 solicitantes de licencia de conducir. En esta prueba se usará como nivel de significancia $\alpha = 0.05$.

En este caso la varianza muestral de las puntuaciones de 30 exámenes fue $s^2 = 162$. El valor del estadístico de prueba chi-cuadrada es el siguiente:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(30-1)(162)}{100} = 46.98$$

Ahora, queda calcular el valor-*p*. Mediante la tabla 11.1 y $n - 1 = 30 - 1 = 29$ grados de libertad (gl) se encuentra lo siguiente.

Área en la cola superior	**0.10**	**0.05**	**0.025**	**0.01**
Valor χ^2 (29 *gl*)	39.087	42.557	45.722	49.588

$\chi^2 = 46.98$ (entre 45.722 y 49.588)

De manera que el valor del estadístico de prueba $\chi^2 = 46.98$ corresponde a un área entre 0.025 y 0.01 en la cola superior de la distribución chi-cuadrada. Al duplicar este valor, se tiene que el

valor-p está entre 0.05 y 0.02. Con Excel o Minitab se encuentra el exacto valor-p = 0.374. Como el valor-$p \leq \alpha = 0.05$, se rechaza H_0 y se concluye que en el nuevo examen las puntuaciones presentan una varianza poblacional distinta a la varianza histórica de $\sigma^2 = 100$. Un resumen de las pruebas de hipótesis para la varianza poblacional se presenta en la tabla 11.2.

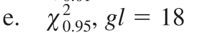
TABLA 11.2 PRUEBAS DE HIPÓTESIS PARA LA VARIANZA POBLACIONAL

	Prueba de la cola inferior	**Prueba de la cola superior**	**Prueba de dos colas**
Hipótesis	$H_0: \sigma^2 \geq \sigma_0^2$ $H_a: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 \leq \sigma_0^2$ $H_a: \sigma^2 > \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$ $H_a: \sigma^2 \neq \sigma_0^2$
Estadístico de prueba	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$
Regla de rechazo: método del valor-*p*	Rechazar H_0 si valor-$p \leq \alpha$	Rechazar H_0 si valor-$p \leq \alpha$	Rechazar H_0 si valor-$p \leq \alpha$
Regla de rechazo: método del valor crítico	Rechazar H_0 si $\chi^2 \leq \chi^2_{(1-\alpha)}$	Rechazar H_0 si $\chi^2 \geq \chi^2_{\alpha}$	Rechazar H_0 si $\chi^2 \leq \chi^2_{(1-\alpha/2)}$ o si $\chi^2 \geq \chi^2_{\alpha/2}$

Ejercicios

Métodos

1. En la tabla 11.1 o en la tabla 3 del apéndice B encuentre los valores siguientes de la distribución chi-cuadrada.
   a. $\chi^2_{0.05}$, $gl = 5$
   b. $\chi^2_{0.025}$, $gl = 15$
   c. $\chi^2_{0.975}$, $gl = 20$
   d. $\chi^2_{0.01}$, $gl = 10$
   e. $\chi^2_{0.95}$, $gl = 18$







2. En una muestra de 20 elementos la desviación estándar muestral es 5.
   a. Calcule una estimación por intervalo de confianza de 90% para la varianza poblacional.
   b. Calcule una estimación por intervalo de confianza de 95% para la varianza poblacional.
   c. Calcule una estimación por intervalo de confianza de 95% para la desviación estándar poblacional.

3. En una muestra de 16 elementos la desviación estándar muestral es 9.5. Pruebe la hipótesis siguiente usando $\alpha = 0.05$. ¿A que conclusión llega? Use tanto el método del valor-p como el del valor crítico.

$$H_0: \sigma^2 \leq 50$$
$$H_a: \sigma^2 > 50$$

Aplicaciones

4. En la industria farmacéutica la varianza en los pesos de los medicamentos es trascendental. Considere un medicamento cuyo peso está dado en gramos y una muestra de 18 unidades de este medicamento, la varianza muestral es $s^2 = 0.36$.
   a. Dé un intervalo de 90% de confianza para estimar la varianza poblacional de los pesos de este medicamento.
   b. Proporcione un intervalo de 90% de confianza para estimar la desviación estándar poblacional.

5. A continuación se presentan los precios de las rentas de un automóvil por día en ocho ciudades.

Ciudad	Renta de un automóvil por día ($)
Atlanta	47
Chicago	50
Dallas	53
New Orleans	45
Phoenix	40
Pittsburgh	43
San Francisco	39
Seattle	37

a. Calcule la varianza y la desviación estándar de estos datos.
b. Dé la estimación por intervalo de confianza de 95% por día para la varianza poblacional de los precios de renta de un automóvil por día.
c. Dé la estimación por intervalo de confianza de 95% para la desviación estándar poblacional.

6. La Fidelity Growth & Income recibe fondos mutualistas de tres estrellas, o neutrales, clasificados por Mornigstar. A continuación se presentan los rendimientos porcentuales trimestrales en el periodo de cinco años que va de 2001 a 2005 (*Mornigstar Funds 500*, 2006).



	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2001	−10.91	5.80	−9.64	6.45
2002	0.83	−10.48	−14.03	5.58
2003	−2.27	10.43	0.85	9.33
2004	1.34	1.11	−0.77	8.03
2005	−2.46	0.89	2.55	1.78

a. Calcule media, varianza y desviación estándar de estos rendimientos trimestrales.
b. Los analistas financieros suelen usar la desviación estándar como una medida del riesgo de acciones y fondos mutualistas. Dé un intervalo de 95% de confianza para la desviación estándar poblacional del rendimiento trimestral de los fondos mutualistas de Fidelity Growth & Income.

7. Para analizar el riesgo o la volatilidad al invertir en las acciones comunes de Chevron Corpotation se toma una muestra de rendimiento porcentual total mensual. A continuación se presentan los rendimientos de los 12 meses de 2005 (*Compustat*, 24 de febrero de 2006). El rendimiento total es el precio más cualquier dividendo pagado.

Mes	Rendimiento (%)	Mes	Rendimiento (%)
Enero	3.60	Julio	3.74
Febrero	14.86	Agosto	6.62
Marzo	−6.07	Septiembre	5.42
Abril	−10.82	Octubre	−11.83
Mayo	4.29	Noviembre	1.21
Junio	3.98	Diciembre	−0.94

a. Calcule la varianza muestral y la desviación estándar muestral como medidas de la volatilidad del rendimiento mensual total de Chevron.
b. Dé un intervalo de 95% de confianza para la varianza poblacional.
c. Proporcione un intervalo de 95% de confianza para la desviación estándar poblacional.

8. Un grupo de 12 analistas de seguridad proporcionó estimaciones, para el año 2001, de las ganancias por acción de Qualcomm, Inc. (*Zacks.com*, 13 de junio de 2000). Los datos son los siguientes:

1.40 1.40 1.45 1.49 1.37 1.27 1.40 1.55 1.40 1.42 1.48 1.63

a. Calcule la varianza muestral de las estimaciones de ganancia por acción.
b. Calcule la desviación estándar muestral de las estimaciones de ganancia por acción.
c. Dé una estimación por intervalo de confianza de 95% para la varianza poblacional y para la desviación estándar poblacional.



9. Una pieza para automóviles debe fabricarse con medidas de tolerancia muy estrechas para que sea aceptada por el cliente. Las especificaciones de producción indican que la varianza máxima en la longitud de la pieza debe ser 0.0004. Suponga que en 30 piezas la varianza muestral encontrada es $s^2 = 0.0005$. Use $\alpha = 0.05$ para probar si se está violando la especificación para la varianza poblacional

10. La desviación estándar promedio del rendimiento anual de fondos mutualistas de acciones de capital grande es 18.2% (*The Top Mutual Funds*, AAII, 2004). La desviación estándar muestral en una muestra de 36 fondos mutualistas Vanguard PRIMECAP es 22%. Realice una prueba de hipótesis para determinar si la desviación estándar de los fondos Vanguard es mayor que la desviación estándar promedio de los fondos mutualistas de capital grande. Con nivel de significancia 0.05, ¿cuál es la conclusión?

11. Las tasas de interés en hipotecas para vivienda a 30 años con plazos fijos varían en Estados Unidos. En el verano de 2000, los datos de varias partes del país indicaban que la desviación estándar de las tasas de interés era 0.096 (*The Wall Street Journal*, 8 de sepembre de 2000). La varianza correspondiente sería $(0.096)^2 = 0.009216$. En un estudio realizado en 2001, las tasas de interés en préstamos a 30 años con plazo fijo en una muestra de 20 instituciones de préstamo mostraron una desviación estándar muestral de 0.114. Realice una prueba de hipótesis usando H_0: $\sigma^2 = 0.009216$ para determinar si los datos muestrales indican que la variabilidad en las tasas de interés ha cambiado. Use $V = 0.05$ y dé la conclusión.

12. En un estudio de *Fortune* se encontró que la varianza en la cantidad de vehículos que poseen o rentan los suscriptores de la revista *Fortune* es 0.94. Suponga que en una muestra de 12 suscriptores de otra revista se encuentran los datos siguientes sobre la cantidad de vehículos que poseen o rentan dichos suscriptores: 2, 1, 2, 0, 3, 2, 2, 1, 2, 1, 0 y 1.
a. Calcule la varianza muestral de la cantidad de vehículos que poseen o rentan estos 12 suscriptores.
b. Pruebe la hipótesis H_0: $\sigma^2 = 0.94$ para determinar si la varianza del número de vehículos que poseen o rentan los suscriptores de la otra revista difiere de la propia de *Fortune*, que es $\sigma^2 = 0.94$. Con un nivel de significancia 0.05, ¿cuál es la conclusión?



Inferencias acerca de dos varianzas poblacionales

En algunas aplicaciones estadísticas interesa comparar las varianzas de las calidades de producto obtenido mediante dos métodos de producción diferentes, o las varianzas de tiempos de fabricación empleando dos métodos diferentes, o las varianzas de las temperaturas que se tienen con dos dispositivos distintos de calentamiento. Para comparar dos varianzas poblacionales, se emplean datos obtenidos de dos muestras aleatorias independientes, una de la población 1 y otra de la población 2. Para hacer las inferencias acerca de las dos varianzas poblacionales s_1^2 y s_2^2 se usan las dos varianzas muestrales σ_1^2 y σ_2^2. Cuando las varianzas de dos poblaciones normales son iguales ($\sigma_1^2 = \sigma_2^2$), la distribución muestral de la proporción entre las dos varianzas muestrales s_1^2/s_2^2 es la siguiente.

DISTRIBUCIÓN MUESTRAL DE s_1^2/s_2^2 CUANDO $\sigma_1^2 = \sigma_2^2$

Cuando se toman muestras aleatorias simples independientes de tamaños n_1 y n_2 de dos poblaciones normales con varianzas iguales, la distribución muestral de

$$\frac{s_1^2}{s_2^2} \tag{11.9}$$

La distribución F se basa en muestras de dos poblaciones normales.

es una distribución F con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador; s_1^2 es la varianza muestral de la muestra aleatoria de n_1 elementos tomados de la población 1, y s_2^2 es la varianza muestral de la muestra aleatoria de n_2 elementos tomados de la población 2.

La figura 11.4 es una gráfica de la distribución F con 20 grados de libertad tanto en el numerador como en el denominador. Como se ve en esta gráfica, la distribución F no es simétrica y los valores F no pueden ser negativos. La forma de cada distribución F depende de los grados de libertad en el numerador y de los grados de libertad en el denominador.

Para denotar el valor F correspondiente a un área o probabilidad α en la cola superior de la distribución, se usa la notación F_α. Por ejemplo, como aparece en la figura 11.4, $F_{0.05}$ corresponde a un área de 0.05 en la cola superior de la distribución F con 20 grados de libertad en el numerador y 20 grados de libertad en el denominador. El valor de $F_{0.05}$ se encuentra en la tabla de la distribución F, parte de la cual se presenta en la tabla 11.3. Usando 20 grados de libertad en el numerador, 20 grados de libertad en el denominador y el renglón correspondiente a un área de 0.05 en la cola superior de la distribución, se encuentra $F_{0.05} = 2.12$. Observe que la tabla sirve para hallar valores F correspondientes a áreas de 0.10, 0.05, 0.025 y 0.01 en la cola superior. En la tabla 4 del apéndice B se encuentra una tabla más completa de la distribución F.

Ahora se verá cómo usar la distribución F para realizar una prueba de hipótesis para las varianzas de dos poblaciones. Se empieza con una prueba sobre la igualdad de las dos varianzas poblacionales. Las hipótesis son las siguientes:

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_a: \sigma_1^2 \neq \sigma_2^2$$

De manera tentativa se supone que las varianzas poblacionales son iguales. Si se rechaza H_0 se llega a la conclusión de que las varianzas poblacionales no son iguales.

Para realizar esta prueba de hipótesis se requieren dos muestras aleatorias independientes, una de cada población. Se calculan las dos varianzas muestrales. A la población en la que se encuentre la *mayor* varianza muestral se le considera la población 1. De manera que el tamaño de muestra n_1 y la varianza muestral s_1^2 corresponden a la población 1 y el tamaño de muestra n_2 y

FIGURA 11.4 DISTRIBUCIÓN F CON 20 GRADOS DE LIBERTAD EN EL NUMERADOR Y 20 GRADOS DE LIBERTAD EN EL DENOMINADOR

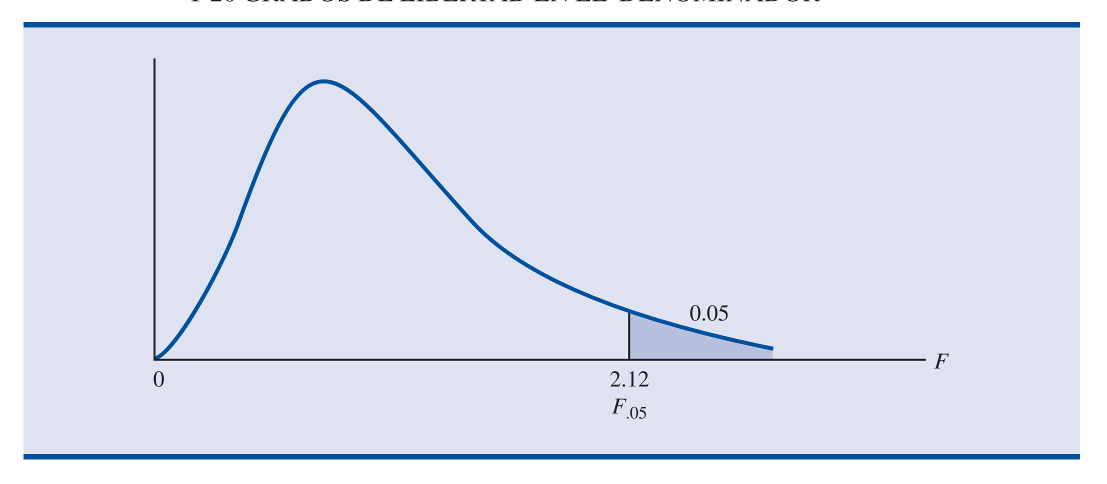
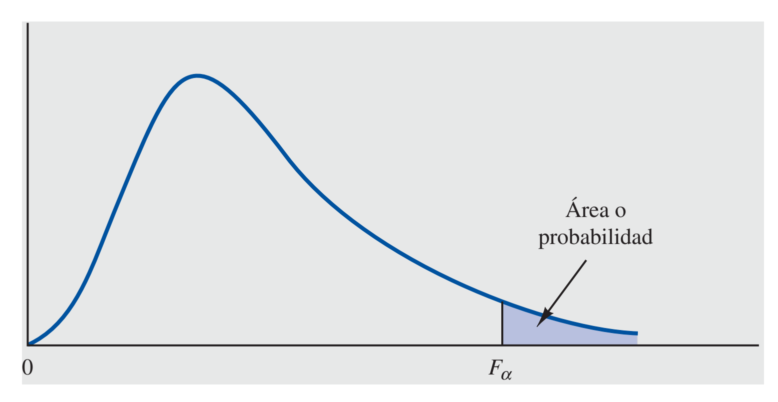


TABLA 11.3 ALGUNOS VALORES DE LA TABLA DE LA DISTRIBUCIÓN F*



Grados de libertad en el denominador	Área en la cola superior	Grados de libertad en el numerador 10	15	20	25	30
10	0.10	2.32	2.24	2.20	2.17	2.16
	0.05	2.98	2.85	2.77	2.73	2.70
	0.025	3.72	3.52	3.42	3.35	3.31
	0.01	4.85	4.56	4.41	4.31	4.25
15	0.10	2.06	1.97	1.92	1.89	1.87
	0.05	2.54	2.40	2.33	2.28	2.25
	0.025	3.06	2.86	2.76	2.69	2.64
	0.01	3.80	3.52	3.37	3.28	3.21
20	0.10	1.94	1.84	1.79	1.76	1.74
	0.05	2.35	2.20	2.12	2.07	2.04
	0.025	2.77	2.57	2.46	2.40	2.35
	0.01	3.37	3.09	2.94	2.84	2.78
25	0.10	1.87	1.77	1.72	1.68	1.66
	0.05	2.24	2.09	2.01	1.96	1.92
	0.025	2.61	2.41	2.30	2.23	2.18
	0.01	3.13	2.85	2.70	2.60	2.54
30	0.10	1.82	1.72	1.67	1.63	1.61
	0.05	2.16	2.01	1.93	1.88	1.84
	0.025	2.51	2.31	2.20	2.12	2.07
	0.01	2.98	2.70	2.55	2.45	2.39

**Nota:* La tabla 4 del apéndice B es una tabla más completa.

la varianza muestral s_2^2 corresponden a la población 2. Con base en la suposición de que las dos poblaciones tengan una distribución normal, la relación entre las varianzas muestrales proporciona el siguiente estadístico de prueba F.

ESTADÍSTICO DE PRUEBA PARA PRUEBAS DE HIPÓTESIS ACERCA DE VARIANZAS POBLACIONALES CON $\sigma_1^2 = \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2} \tag{11.10}$$

Al denotar como población 1 a la población que tiene mayor varianza muestral, el estadístico de prueba tiene una distribución F con $n_1 - 1$ grados de libertad en el numerador y con $n_2 - 1$ grados de libertad en el denominador.

Como el estadístico de prueba F se construye con la varianza muestral más grande s_1^2 en el numerador, el valor del estadístico de prueba se encontrará siempre en la cola superior de la distribución F. Por tanto, las tablas de la distribución F, como la presentada en la tabla 11.3 y en la tabla 4 del apéndice B, únicamente necesitan proporcionar áreas o probabilidades en la cola superior. Si no se construyera de este modo el estadístico de prueba, serían necesarias áreas o probabilidades en la cola inferior. En tal caso se necesitarían más cálculos o tablas más extensas para la distribución F. A continuación se presenta un ejemplo de una prueba de hipótesis para la igualdad de dos varianzas poblacionales.

Dullus County Schools está por renovar el contrato del servicio de autobús para el año entrante y debe decidirse entre dos empresas que prestan el servicio, la empresa Milbank y la empresa Gulf Park. Como medida de la calidad del servicio se emplea la varianza en los tiempos en que llega a recoger/dejar a las personas. Poca varianza indica un mejor servicio, un servicio de mayor calidad. Si las varianzas de las dos empresas son iguales, la escuela elegirá la empresa que ofrezca mejores condiciones financieras. Pero si hay una diferencia significativa en las varianzas, la escuela preferirá la empresa con la menor varianza o mejor servicio. Las hipótesis en este caso son las siguientes:

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_a: \sigma_1^2 \neq \sigma_2^2$$

Si se rechaza H_0, se concluirá que los servicios no tienen la misma calidad. Para realizar esta prueba de hipótesis se usa como nivel de significancia $\alpha = 0.10$

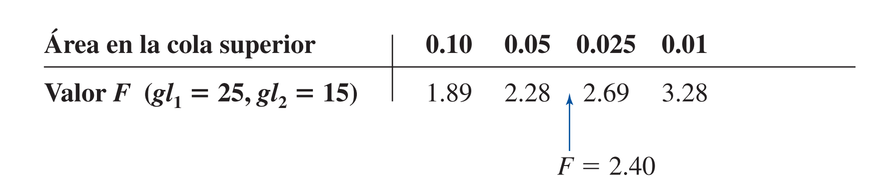


En una muestra de 26 tiempos de llegada de la empresa Milbank la varianza muestral es 48 y en una muestra de 16 tiempos de llegada de la empresa Gulf Park la varianza muestral es 20. Como la varianza en la muestra de Milbank es la mayor, Milbank será la población 1. Usando la ecuación (11.10) se encuentra el valor del estadístico de prueba:

$$F = \frac{s_1^2}{s_2^2} = \frac{48}{20} = 2.40$$

La distribución F es la que tiene $n_1 - 1 = 26 - 1 = 25$ grados de libertad en el numerador y $n_2 - 1 = 16 - 1 = 15$ grados de libertad en el denominador.

Igual que en las otras pruebas de hipótesis, para llegar a una conclusión en esta prueba de hipótesis se puede emplear el método del valor-p o el método del valor crítico. En la tabla 11.3 se encuentran las siguientes áreas en la cola superior correspondientes a los valores F de una distribución F con 25 grados de libertad en el numerador y 15 grados de libertad en el denominador.



Área en la cola superior	**0.10**	**0.05**	**0.025**	**0.01**
Valor F ($gl_1 = 25, gl_2 = 15$)	1.89	2.28	2.69	3.28

Como $F = 2.40$ está entre 2.28 y 2.69, el área en la cola superior de la distribución está entre 0.05 y 0.025. Como se trata de una prueba de dos colas, se duplica el área de la cola superior, y

se obtiene un valor-p entre 0.10 y 0.05. Como se eligió $\alpha = 0.10$ como nivel de significancia, el valor-$p < \alpha = 0.10$. Por tanto, se rechaza la hipótesis nula. Esto lleva a la conclusión de que los dos servicios de autobús difieren en términos de la varianza de los tiempos en que llegan a recoger/dejar a las personas. Se le recomienda a la escuela el servicio de menor varianza o el servicio mejor que es el ofrecido por la empresa Gulf Park.

Usando Excel o Minitab se encuentra que el estadístico de prueba $F = 2.40$, corresponde a un valor-$p = 0.0811$. Como $0.0811 < \alpha = 0.10$, se rechaza la hipótesis nula de que las dos varianzas poblacionales son iguales.

Para usar el método del valor crítico en una prueba de hipótesis de dos colas con $\alpha = 0.10$, se toman los valores críticos correspondientes a un área $\alpha/2 = 0.10/2 = 0.05$ en cada cola de la distribución. Como el valor del estadístico de prueba calculado con la ecuación (11.10) está siempre en la cola superior, basta determinar el valor crítico en la cola superior. En la tabla 11.3 se encuentra que $F_{0.05} = 2.28$. Así, aun cuando se trata de una prueba de dos colas, la regla de rechazo es la siguiente:

$$\text{Rechazar } H_0 \text{ si } F \geq 2.28$$

Como el estadístico de prueba es $F = 2.40$, mayor que 2.28, se rechaza H_0 y se concluye que los dos servicios difieren en términos de la varianza en los tiempos en que llegan a recoger/dejar a las personas.

También se pueden realizar pruebas de una cola para dos varianzas poblacionales. En estos casos se usa la distribución F para determinar si una varianza poblacional es significativamente mayor que la otra. Una prueba de hipótesis para dos varianzas poblacionales se formula siempre como una prueba de la *cola superior:*

Una prueba de hipótesis de una cola para dos varianzas poblacionales siempre se formula como una prueba de la cola superior. Esto elimina la necesidad de tener valores F de la cola inferior.

$$H_0\colon \sigma_1^2 \leq \sigma_2^2$$
$$H_a\colon \sigma_1^2 > \sigma_2^2$$

En esta forma de una prueba de hipótesis, el valor-p y el valor crítico siempre se encuentran en la cola superior de la distribución F. De esta manera, sólo se necesitan los valores F de la cola superior, lo cual simplifica tanto los cálculos como la tabla de la distribución F.

A continuación se ilustra el uso de la distribución F para realizar una prueba de una cola para las varianzas de dos poblaciones, empleando una encuesta sobre opinión pública. Para estudiar las actitudes frente a los asuntos políticos actuales se emplea una muestra de 31 hombres y otra de 41 mujeres. Al investigador que realiza el estudio le interesa saber si los datos muestrales indican que entre las mujeres hay mayor variación en las actitudes respecto de los asuntos políticos que entre los hombres. En la forma de la prueba de hipótesis dada arriba, las mujeres serán la población 1 y los hombres la población 2. La prueba de hipótesis se plantea como sigue:

$$H_0\colon \sigma_{\text{mujeres}}^2 \leq \sigma_{\text{hombres}}^2$$
$$H_a\colon \sigma_{\text{mujeres}}^2 > \sigma_{\text{hombres}}^2$$

Rechazar H_0 dará al investigador el respaldo estadístico necesario para concluir que las mujeres muestran mayor variación en las actitudes respecto a los asuntos políticos.

Con la varianza muestral de las mujeres en el numerador y la varianza muestral de los hombres en el denominador, la distribución F tendrá $n_1 - 1 = 41 - 1 = 40$ grados de libertad en el numerador y $n_2 - 1 = 31 - 1 = 30$ grados de libertad en el denominador. En esta prueba de hipótesis se usa $\alpha = 0.05$ como nivel de significancia. Como resultado de la prueba se encontró una varianza muestral para las mujeres $s_1^2 = 120$ y una varianza muestral para los hombres $s_2^2 = 80$. El estadístico de prueba es el siguiente.

$$F = \frac{s_1^2}{s_2^2} = \frac{120}{80} = 1.50$$

TABLA 11.4 RESUMEN DE LAS PRUEBAS DE HIPÓTESIS ACERCA DE DOS VARIANZAS POBLACIONALES

	Prueba de la cola superior	Prueba de la cola inferior
Hipótesis	$H_0: \sigma_1^2 \leq \sigma_2^2$ $H_a: \sigma_1^2 > \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_a: \sigma_1^2 \neq \sigma_2^2$ Nota: La población 1 tiene la varianza muestral más grande
Estadístico de prueba	$F = \dfrac{s_1^2}{s_2^2}$	$F = \dfrac{s_1^2}{s_2^2}$
Regla de rechazo: método del valor-*p*	Rechazar H_0 si valor-$p \leq \alpha$	Rechazar H_0 si valor-$p \leq \alpha$
Regla de rechazo: método del valor crítico	Rechazar H_0 si $F \geq F_\alpha$	Rechazar H_0 si $F \geq F_{\alpha/2}$

En la tabla 4 del apéndice B la distribución F con 40 grados de libertad en el numerador y 30 grados de libertad en el denominador da $F_{0.10} = 1.57$. Como el estadístico de prueba, $F = 1.50$, es menor que 1.57, el área en la cola superior debe ser mayor que 0.10. Por ende, el valor-p es mayor a 0.10. Usando Excel o Minitab se encuentra que el valor-$p = 0.1256$. Como el valor-p $> \alpha = 0.05$, no se puede rechazar H_0. Por tanto, los resultados muestrales no favorecen la conclusión de que entre las mujeres haya mayor variación en la actitud frente a los asuntos políticos que entre los hombres. En la tabla 11.4 se presenta un resumen de las pruebas de hipótesis para dos varianzas poblacionales.

NOTAS Y COMENTARIOS

Las investigaciones confirman el hecho de que para usar la distribución F es importante suponer que las poblaciones tienen una distribución normal. La distribución F no se puede usar a menos que sea razonable suponer que ambas poblaciones tienen una distribución por lo menos aproximadamente normal.

Ejercicios

Métodos

13. En la tabla 4 del apéndice B halle los valores siguientes de la distribución F.
    a. $F_{0.05}$ con 5 y 10 grados de libertad
    b. $F_{0.025}$ con 20 y 15 grados de libertad
    c. $F_{0.01}$ con 8 y 12 grados de libertad
    d. $F_{0.10}$ con 10 y 20 grados de libertad
14. En una muestra de 16 elementos de la población 1 la varianza muestral es $s_1{}^2 = 5.8$ y en una muestra de 21 elementos de la población 2 la varianza muestral es $s_2^2 = 2.4$. Pruebe las hipótesis siguientes usando 0.05 como nivel de significancia

$$H_0: \sigma_1^2 \leq \sigma_2^2$$
$$H_a: \sigma_1^2 > \sigma_2^2$$

    a. Dé la conclusión a la que se llega usando el método del valor-p.
    b. Repita la prueba usando el método del valor crítico.



15. Considere la prueba de hipótesis siguiente

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_a: \sigma_1^2 \neq \sigma_2^2$$

a. ¿A qué conclusión se llega si $n_1 = 21$, $s_1^2 = 8.2$, $n_2 = 26$ y $s_2^2 = 4.0$? Use $\alpha = 0.05$ y el método del valor-p.
b. Repita la prueba usando el método del valor crítico.

Aplicaciones

16. Media Metrix and Jupiter Communications recogieron datos sobre la cantidad de tiempo que pasan conectados a Internet, por mes, adultos y jóvenes (*USA Today*, 14 de septiembre de 2000). Se concluyó que, en promedio, los adultos pasan más tiempo conectados a Internet que los jóvenes. Suponga que para confirmar esto se realiza otro estudio para el que se toma una muestra de 26 adultos y otra de 30 jóvenes. Las desviaciones estándar de las cantidades de tiempo que pasan conectados a Internet son 94 y 58 minutos, respectivamente. ¿Estos resultados muestrales favorecen la conclusión de que en el caso de los adultos la varianza del tiempo que pasan conectados a Internet es mayor que en el caso de los jóvenes? Use $\alpha = 0.01$. ¿Cuál es el valor-p?

Autoexamen

17. La mayor parte de los individuos saben que el gasto anual medio en reparaciones de un automóvil depende de la antigüedad del automóvil. Un investigador desea saber si la varianza de los gastos anuales que se hacen en reparación también aumenta con la antigüedad del automóvil. En una muestra de 26 automóviles de 4 años de antigüedad la desviación estándar muestral en los gastos anuales en reparación fue \$170 y en una muestra de 25 automóviles de 2 años de antigüedad la desviación estándar muestral en los gastos anuales en reparación fue \$100.
   a. Dé las hipótesis nula y alternativa para la investigación de que la varianza en los gastos anuales de reparación es mayor entre más viejos son los automóviles
   b. Empleando 0.01 como nivel de significancia, ¿cuál es la conclusión? ¿Cuál es el valor-p? Analice lo razonable de sus hallazgos.

18. En 10 empresas de la industria aérea la desviación estándar en las ganancias a 12 meses por acción fue 4.27 y en 7 empresas de la industria automotriz la desviación estándar en las ganancias a 12 meses por acción fue 2.27 (*BusinessWeek*, 14 de agosto de 2000). Realice una prueba para varianzas iguales con $\alpha = 0.05$. ¿Cuál es la conclusión acerca de la variabilidad de las ganancias por acción en la industria aérea y en la industria automotriz?

19. La varianza en un proceso de producción es un indicador importante de la calidad del proceso. Las varianzas grandes representan una oportunidad para mejorar un proceso, hallando maneras de reducir esa varianza. Realice una prueba estadística para determinar si existe una diferencia significativa entre las varianzas de los pesos de las bolsas procesadas con dos máquinas diferentes. Use 0.05 como nivel de significancia. ¿Cuál es la conclusión? ¿Representa alguna de las dos máquinas una oportunidad para mejorar la calidad?



Bags

Máquina 1	2.95	3.45	3.50	3.75	3.48	3.26	3.33	3.20
	3.16	3.20	3.22	3.38	3.90	3.36	3.25	3.28
	3.20	3.22	2.98	3.45	3.70	3.34	3.18	3.35
	3.12							
Máquina 2	3.22	3.30	3.34	3.28	3.29	3.25	3.30	3.27
	3.38	3.34	3.35	3.19	3.35	3.05	3.36	3.28
	3.30	3.28	3.30	3.20	3.16	3.33		

20. De acuerdo con datos obtenidos en un estudio, en las empresas de contadores públicos la varianza de los salarios anuales de los empleados de mayor antigüedad es aproximadamente 2.1 y la varianza de los salarios anuales de los gerentes es alrededor de 11.1. Estos datos están dados en miles de dólares. Si estos datos se obtuvieron de muestras de 25 empleados de mayor antigüedad y 26 gerentes, pruebe la hipótesis de que las varianzas poblacionales de estos dos salarios son iguales. Con 0.05 como nivel de significancia, ¿cuál es su conclusión?

21. Fidelity Magellan es un fondo mutualista de capital grande y Fidelity Small Cap Stock es un fondo mutualista de capital pequeño (*Morningstar Funds 500*, 2006). La desviación estándar de ambos fondos se calculó empleando muestras aleatorias de tamaño 26. La desviación estándar muestral de Fidelity Magellan fue 8.89% y la desviación estándar muestral de Fidelity Small Cap Stock fue 13.03%. Los analistas financieros suelen usar la desviación estándar como una medida del riesgo. Realice una prueba de hipótesis para determinar si los fondos de capital pequeño son más riesgosos que los fondos de capital grande. Use $\alpha = 0.05$ como nivel de significancia.

22. Una hipótesis de investigación sostiene que la varianza de las distancias de frenado de los automóviles sobre pavimento húmedo es mayor que la varianza de las distancias de frenado de los automóviles sobre pavimento seco. En un estudio a 16 automóviles que iban a una misma velocidad se les hizo frenar sobre pavimento húmedo y después sobre pavimento seco. En pavimento húmedo la desviación estándar de las distancias de frenado fue 32 pies. Sobre pavimento seco la desviación estándar es 16 pies.
    a. Con 0.05 como nivel de significancia, ¿los datos muestrales justifican la conclusión de que en las distancias de frenado sobre pavimento húmedo la varianza es mayor que sobre pavimento seco? ¿Cuál es el valor-*p*?
    b. ¿Qué significan las conclusiones estadísticas de este estudio en términos de las recomendaciones para la seguridad al manejar?

Resumen

En este capítulo se presentaron los procedimientos estadísticos que se usan en las inferencias acerca de varianzas poblacionales. Se introdujeron dos distribuciones de probabilidad nuevas: la distribución chi-cuadrada y la distribución *F*. La distribución chi-cuadrada se usa en estimación por intervalos y pruebas de hipótesis para la varianza de una población normal.

Se ilustró el uso de la distribución *F* en pruebas de hipótesis para las varianzas de dos poblaciones normales. En particular, se mostró que si se tienen muestras aleatorias simples independientes de tamaños n_1 y n_2, tomadas de dos poblaciones normales que tienen varianzas iguales $\sigma_1^2 = \sigma_2^2$, la distribución muestral de la razón entre las dos varianzas muestrales s_1^2/s_2^2 tiene una distribución *F* con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador.

Fórmulas clave

Estimación por intervalo para una varianza poblacional

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}} \quad \textbf{(11.7)}$$

Estadístico de prueba en una prueba de hipótesis para la varianza poblacional

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \quad \textbf{(11.8)}$$

Estadístico de prueba en una prueba de hipótesis para varianzas poblacionales con $\sigma_1^2 = \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2} \quad \textbf{(11.10)}$$

Ejercicios complementarios

23. Por cuestiones de personal, los administradores de un hotel desean conocer la variabilidad en la cantidad de habitaciones ocupadas por día en una determinada temporada del año. En una muestra de 20 días la media muestral es 290 habitaciones ocupadas por día y la desviación estándar es 30 habitaciones.
    a. Dé la estimación puntual de la varianza poblacional.
    b. Dé una estimación por intervalo de 90% de confianza para la varianza poblacional.
    c. Proporcione una estimación por intervalo de 90% de confianza para la desviación estándar poblacional.
24. Las ofertas públicas iniciales (OPI) de acciones suelen estar subvaluadas. La desviación estándar mide la dispersión o variación del indicador subvaluación-sobrevaluación. En una muestra de 13 OPI canadienses, que fueron después negociadas en la bolsa de cambio de Toronto, esta desviación estándar fue de 14.95. Dé una estimación por intervalo de confianza de 95% para la desviación estándar poblacional del indicador subvaluación-sobrevaluación.
25. A continuación se presentan los costos estimados de manutención por día de un ejecutivo que viaja a varias ciudades importantes. Las estimaciones comprenden una habitación individual en un hotel de cuatro estrellas, bebidas, desayuno, taxis y costos incidentales.

archivo en CD
Travel

Ciudad	Costo de manutención por día	Ciudad	Costo de manutención por día
Bangkok	$242.87	Madrid	$283.56
Bogotá	260.93	Mexico City	212.00
Bombay	139.16	Milan	284.08
Cairo	194.19	Paris	436.72
Dublín	260.76	Rio de Janeiro	240.87
Frankfurt	355.36	Seoul	310.41
Hong Kong	346.32	Tel Aviv	223.73
Johannesburg	165.37	Toronto	181.25
Lima	250.08	Warsaw	238.20
Londres	326.76	Washington, D.C.	250.61

    a. Calcule la media muestral.
    b. Calcule la desviación estándar muestral.
    c. Calcule un intervalo de 95% de confianza para la desviación estándar poblacional.
26. La variabilidad es crucial en la fabricación de cojinetes de bolas. Una varianza grande en el tamaño de estos cojinetes ocasiona que no trabajen bien y que se desgasten rápidamente. Las normas de producción exigen una varianza máxima de 0.0001 en la medida de los cojinetes dada en pulgadas. En una muestra de 15 cojinetes, la desviación estándar muestral fue 0.014 pulgadas.
    a. Use $\alpha = 0.10$ para determinar si la muestra indica que se ha excedido la varianza máxima indicada.
    b. Dé una estimación por intervalo de confianza de 90% para la varianza poblacional de los cojinetes.
27. La varianza en el llenado de las cajas de cereal debe ser 0.02 o menos. En una muestra de 41 cajas de cereal la desviación estándar muestral es 0.16 onzas. Use $\alpha = 0.05$ para determinar si la varianza en el llenado de las cajas de cereal ha excedido la especificación.
28. Una empresa de transporte de carga asegura tiempos uniformes en sus entregas. En una muestra de 22 entregas la varianza muestral fue 1.5. Realice una prueba de hipótesis para determinar si se puede rechazar H_0: $\sigma^2 \leq 1$. Use $\alpha = 0.10$.
29. En una muestra de 9 días de los últimos seis meses se encontró que un dentista había tratado los siguientes números de pacientes: 22, 25, 20, 18, 15, 22, 24, 19 y 26. Si el número de pacientes atendidos por día tiene una distribución normal, ¿un análisis de estos datos muestrales permitiría

rechazar la hipótesis de que la varianza de la cantidad de pacientes atendidos por día es 10? Use un nivel de significancia de 0.10. ¿Cuál es la conclusión?

30. La desviación estándar muestral del número de pasajeros que toman un determinado vuelo de una línea aérea es 8. Una estimación por intervalo de confianza de 95% para la desviación estándar poblacional es el que va de 5.86 pasajeros a 12.62 pasajeros.
    a. ¿El tamaño de la muestra usado en este análisis estadístico fue 10 o 15?
    b. Suponga que la desviación estándar muestral $s = 8$ se obtuvo de una muestra de 25 vuelos. ¿Cuál sería el cambio en el intervalo de confianza para la desviación estándar poblacional? Calcule una estimación por intervalo de confianza de 95% para s con un tamaño de muestra de 25.

31. En los principales mercados de acciones diario existe un grupo de principales ganadoras en precio (acciones que registran las mayores alzas). Un día la desviación estándar del cambio porcentual en una muestra de 10 de las principales ganadoras que forman parte del NASDAQ fue 15.8. Ese mismo día la desviación estándar del cambio porcentual en una muestra de 10 de las principales ganadoras que forman parte del NYSE fue 7.9 (*USA Today*, 14 de septiembre de 2000). Realice una prueba para varianzas poblacionales iguales para saber si es posible concluir que existe diferencia en la volatilidad de las ganadoras principales de los dos grupos. Use $\alpha = 0.10$. ¿Cuál es la conclusión?

32. En los promedios de calificaciones de 352 estudiantes que terminaron un curso de contabilidad financiera, la desviación estándar es 0.940. En los promedios de calificaciones de 73 estudiantes que no aprobaron el mismo curso la desviación estándar es 0.797. ¿Estos datos indican alguna diferencia entre las varianzas de los promedios de las calificaciones de quienes terminaron el curso y de los que no lo aprobaron? Use 0.05 como nivel de significancia. *Nota:* $F_{0.025}$ con 351 y 72 grados de libertad es 1.466.

33. El departamento de contabilidad analiza la varianza de los costos unitarios semanales en los informes de dos departamentos de producción. En una muestra de 16 informes de costos de cada uno de los departamentos, las varianzas de los costos fueron 2.3 y 5.4, respectivamente. ¿La muestra es suficiente para concluir que los dos departamentos difieren en términos de la varianza en los costos unitarios? Use $\alpha = 0.10$

34. Al probar dos métodos de fabricación se da el tiempo requerido por cada uno de ellos. Use $\alpha =$ 0.10 para probar la igualdad de las dos varianza poblacionales.

	Método A	**Método B**
Tamaño de la muestra	$n_1 = 31$	$n_2 = 25$
Variación muestral	$s_1^2 = 25$	$s_2^2 = 12$

Caso problema Programa de capacitación para la Fuerza Aérea

En la fuerza aérea, en un curso introductorio sobre electrónica, se emplea un sistema personalizado en el que cada estudiante ve una clase grabada en una videocinta y después se le da un texto de enseñanza programada. Los estudiantes trabajan con el texto en forma independiente hasta que terminan y aprueban un examen. Aquí preocupan los diferentes ritmos en que los estudiantes realizan esta etapa de su capacitación. Algunos asimilan el texto de enseñanza programada relativamente pronto, mientras que otros necesitan mucho más tiempo. Entonces los primeros deben esperar hasta que los estudiantes más lentos estén listos y todo el grupo pueda pasar a otra etapa de la capacitación.

Se ha propuesto un sistema alternativo en el que se emplea enseñanza asistida por computadora. Este método consiste en que todos los estudiantes vean la misma clase grabada y después a cada uno se le asigne una terminal de computadora para continuar con la capacitación. La computadora guía al estudiante, quien trabaja en forma independiente, a través de esta parte de la capacitación.

Para comparar estos dos métodos, el propuesto y el actual, a los integrantes de un nuevo grupo de 122 estudiantes se les asignó en forma aleatoria uno de los métodos de capacitación. Un grupo de 61 estudiantes usó el método del texto programado y el otro grupo de 61 estudiantes usó el método de enseñanza asistida por computadora. Se registró el tiempo, en horas, que necesitó cada estudiante. Los datos que se presentan a continuación se encuentran en el archivo Training del disco compacto que se distribuye con el libro.

Horas necesarias para terminar el curso empleando el método actual

76	76	77	74	76	74	74	77	72	78	73
78	75	80	79	72	69	79	72	70	70	81
76	78	72	82	72	73	71	70	77	78	73
79	82	65	77	79	73	76	81	69	75	75
77	79	76	78	76	76	73	77	84	74	74
69	79	66	70	74	72					



Horas necesarias para terminar el curso empleando el método de enseñanza asistida por computadora

74	75	77	78	74	80	73	73	78	76	76
74	77	69	76	75	72	75	72	76	72	77
73	77	69	77	75	76	74	77	75	78	72
77	78	78	76	75	76	76	75	76	80	77
76	75	73	77	77	77	79	75	75	72	82
76	76	74	72	78	71					

Informe administrativo

1. Use la estadística descriptiva adecuada para resumir las horas que se necesitaron con cada método. ¿Qué semejanzas y diferencias observa entre estos datos muestrales?
2. Utilice los métodos del capítulo 10 para comentar las diferencias entre las medias poblacionales de los dos métodos.
3. Calcule la desviación estándar y la varianza de los datos obtenidos con cada método. Realice una prueba de hipótesis para la igualdad de las varianzas poblacionales en los datos obtenidos con los dos métodos. Explique sus hallazgos.
4. ¿Qué conclusión obtiene acerca de las diferencias entre los dos métodos? ¿Qué recomienda? Explique.
5. ¿Sugiere otros datos o pruebas que sean de utilidad, antes de decidir qué programa de capacitación usar?

Apéndice 11.1 Varianzas poblacionales con Minitab

Aquí se describe cómo usar Minitab para realizar una prueba de hipótesis para dos varianzas poblacionales.



Use los datos de la sección 11.2 del estudio para la elección del servicio de autobús escolar. Los tiempos correspondientes a la empresa Milbank se encuentran en la columna C1 y los tiempos correspondientes a la empresa Gulf Park en la columna C2. Para realizar la prueba de hipótesis H_0: $\sigma_1^2 = \sigma_2^2$ y H_a: $\sigma_1^2 \neq \sigma_2^2$, se sigue el procedimiento que se describe a continuación.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Basic Statistics**

Paso 3. Elegir **2-Variances**
Paso 4. Cuando aparezca el cuadro de diálogo 2-Variances:
Seleccionar **Samples in different columns**
Ingresar C1 en el cuadro **First**
Ingresar C2 en el cuadro **Second**
Clic en **OK**

En la zona con el título F-Test (normal distribution) se desplegará la información sobre la prueba, dando el estadístico de prueba $F = 2.40$ y el valor-$p = 0.81$. Con este procedimiento de Minitab se realiza una prueba de dos colas para la igualdad de las varianzas poblacionales. Por tanto, cuando se usa esta rutina de Minitab para una prueba de una cola, debe recordar que el área en una cola es la mitad del área del valor-p para dos colas, entonces será fácil calcular el valor-p para la prueba de una cola.

Apéndice 11.2 Varianzas poblacionales con Excel

Aquí se describe cómo usar Excel para realizar una prueba de hipótesis para dos varianzas poblacionales.



SchoolBus

Use los datos de la sección 11.2 del estudio para la elección del servicio de autobús escolar. En la hoja de cálculo de Excel aparece en la celda A1 el rótulo Milbank y el rótulo Gulf Park en la celda B1. Los datos muestrales de Milbank se encuentran en las celdas A2:A27 y los datos muestrales de Gulf Park se encuentran en las celdas B2:B17 Los pasos para realizar la prueba de hipótesis H_0: $\sigma_1^2 = \sigma_2^2$ y H_a: $\sigma_1^2 \neq \sigma_2^2$ se presentan a continuación:

Paso 1. Seleccionar el menú **Herramientas**
Paso 2. Elegir **Análisis de datos**
Paso 3. Cuando aparezca el cuadro de diálogo Análisis de datos:
Elegir **Prueba F para varianza de dos muestras**
Clic en **Aceptar**
Paso 4. Cuando aparezca el cuadro de diálogo Prueba F para varianza de dos muestras:
Ingresar A1:A27 en el cuadro **Rango para la variable 1**
Ingresar B1:B17 en el cuadro **Rango para la variable 2**
Seleccionar **Rótulos**
Ingresar 0.05 en cuadro **Alfa**
(*Nota:* En este procedimiento Excel usa alfa como área en la cola superior.)
Seleccionar **Rango de salida** e ingresar C1 en el cuadro correspondiente
Clic en **Aceptar**

En los resultados, en la celda $P(F \leq f) = 0.0405$ se da el área en una cola correspondiente al estadístico de prueba $F = 2.40$. Por tanto, el valor-p para dos colas es $2(0.0405) = 0.081$. Si se trata de una prueba de hipótesis de una cola, el área en una cola que aparece en la celda rotulada $P(F \leq f)$ proporciona la información necesaria para determinar el valor-p de la prueba.

CAPÍTULO 12



Pruebas de bondad de ajuste e independencia

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

UNITED WAY*
ROCHESTER, NUEVA YORK

United Way of Greater Rochester es una organización no lucrativa que se dedica a mejorar la calidad de vida de todas las personas en las siete zonas en las que proporcionan servicios para cubrir las necesidades humanas más importantes de las comunidades.

La campaña anual de United Way/Cruz Roja para la recolección de fondos, que se realiza todas las primaveras, patrocina cientos de programas ofrecidos por más de 200 proveedores de servicios. Estos proveedores atienden una amplia variedad de necesidades humanas —físicas, mentales y sociales— atendiendo a personas de cualquier edad, origen y situación económica

Debido a la gran cantidad de voluntarios, United Way mantiene sus costos de operación a sólo 8 centavos por dólar recaudado.

United Way of Great Rochester decidió hacer un estudio para saber más acerca de la percepción de la comunidad sobre la caridad. Entrevistas enfocadas a grupos fueron realizadas con profesionales, personal de servicio y trabajadores generales para obtener información preliminar sobre sus percepciones. La información obtenida se usó para elaborar los cuestionarios para el estudio. El cuestionario fue probado, modificado y distribuido a 440 personas; se obtuvieron 323 cuestionarios contestados.

A partir de los datos recolectados se consiguieron diversos estadísticos descriptivos, como distribuciones de frecuencias y tabulaciones cruzadas. Una parte importante del análisis fue el uso de tablas de contingencia y de pruebas chi-cuadrada de independencia. Uno de los usos de dichos estadísticos fue determinar si las percepciones sobre los gastos administrativos eran independientes de la ocupación.

Las hipótesis para la prueba de independencia fueron:

H_0: La percepción sobre los gastos administrativos de United Way es independiente de la ocupación del entrevistado.

H_a: La percepción sobre los gastos administrativos de United Way no es independiente de la ocupación del entrevistado

*Los autores agradecen al doctor Philip R. Tyler, consultor de marketing de United Way por proporcionar este artículo para *La estadística en la práctica.*



Los programas de United Way cubren necesidades tanto de los niños como de los adultos. © Ed Bock/CORBIS.

Dos de las preguntas del estudio suministraron los datos para la prueba estadística. Con una de las preguntas se obtenía información sobre las percepciones de los recursos que se destinaban a gastos administrativos (hasta 10%, 11-20% y 21% o más). Con la otra se preguntaba sobre la ocupación del entrevistado.

La prueba chi cuadrada con 0.05 como nivel de significancia llevó a rechazar la hipótesis nula de independencia y, de esta manera, a la conclusión de que las percepciones sobre los gastos administrativos variaban de acuerdo con la ocupación. En realidad los gastos administrativos eran menores que 9%, pero 35% de los entrevistados tenía la percepción de que eran 21% o más. Así que muchos tenían una percepción inadecuada sobre los gastos administrativos. De este grupo, los empleados de líneas de producción, los empleados de oficina, los vendedores y los técnicos profesionistas tenían percepciones más equivocadas que otros grupos.

El estudio sobre la percepción de la comunidad sirvió para que United Way of Rochester hiciera ajustes a sus programas y a sus actividades de recaudación de fondos. En este capítulo verá cómo se realiza una prueba estadística de independencia, como la descrita aquí.

En el capítulo 11 se vio la forma de usar la distribución chi-cuadrada en estimaciones y en pruebas de hipótesis para la varianza poblacional. En el capítulo 12 se presentan otras dos pruebas de hipótesis, ambas establecidas en el uso de la distribución chi-cuadrada. Como en otras pruebas de hipótesis, en éstas se comparan los resultados muestrales con los esperados si la hipótesis nula es verdadera. La conclusión de la prueba de hipótesis se basa en qué tan “cerca” se encuentran los resultados muestrales de los resultados esperados.

En la sección siguiente se presenta la prueba de bondad de ajuste para una población multinomial. Más adelante se ve la prueba de independencia usando tablas de contingencia y después pruebas de bondad de ajuste para distribuciones normales y de Poisson.



12.1 Prueba de bondad de ajuste: una población multinomial

En esta sección se estudia el caso en que cada elemento de una población corresponde a una y sólo a una de varias clases o categorías. A estas poblaciones se les conoce como **poblaciones multinomiales**. La distribución multinomial se puede entender como una extensión de la distribución binomial al caso en el que hay tres o más categorías de resultados. En cada ensayo de un experimento multinomial uno y sólo uno de los resultados ocurre. Se supone que cada ensayo del experimento es independiente y que en todos los ensayos las probabilidades para los resultados permanecen constantes.

Las suposiciones en un experimento multinomial son las mismas que en un experimento binomial, salvo que en el experimento multinomial en cada ensayo hay tres o más resultados.

Como ejemplo, considere un estudio sobre participación en el mercado realizado por la empresa Scott Marketing Research. A lo largo de los años las participaciones en el mercado se han estabilizado en 30% para la empresa A, 50% para la empresa B y 20% para la empresa C. Recién la empresa C ha elaborado un nuevo y mejorado producto para sustituir a uno de sus productos en el mercado y pidió a la empresa Scott Marketing Research que determinara si el nuevo producto modificaría su participación en el mercado.

En este caso, la población de interés es multinomial, cada cliente se clasifica como cliente de la empresa A, de la empresa B o de la empresa C. De manera que se tiene una población multinomial con tres resultados. Para las proporciones se usa la notación siguiente.

$$p_A = \text{participación en el mercado de la empresa A}$$
$$p_B = \text{participación en el mercado de la empresa B}$$
$$p_C = \text{participación en el mercado de la empresa C}$$

Scott Marketing Research realizará un estudio muestral y calculará la proporción que prefiere el producto de cada empresa. Después aplicará una prueba de hipótesis para ver si el nuevo producto modifica las participaciones en el mercado. Suponga que el nuevo producto de la empresa C no modifica las participaciones en el mercado; entonces, las hipótesis nula y alternativa serán las siguientes.

$$H_0\text{: } p_A = 0.30, p_B = 0.50, \text{ y } p_C = 0.20$$
$$H_a\text{: Las proporciones poblacionales no son}$$
$$p_A = 0.30, p_B = 0.50, \text{ y } p_C = 0.20$$

Si los resultados muestrales llevan al rechazo de H_0 Scott Marketing Research tendrá evidencias de que la introducción del nuevo producto afecta las participaciones del mercado.

Considere que para este estudio la empresa de investigación de mercado ha empleado un panel de 200 consumidores. A cada individuo se le pide que indique su preferencia entre el producto de la empresa A, el producto de la empresa B o el nuevo producto de la empresa C. Las 200 respuestas obtenidas se presentan a continuación en forma resumida.

El panel de 200 consumidores en donde a cada consumidor se le pide que elija una de tres alternativas, es equivalente a un experimento multinomial consistente de 200 ensayos.

Frecuencia esperada		
Producto de la empresa A	**Producto de la empresa B**	**Producto de la empresa C**
48	98	54

Ahora se realiza la **prueba de bondad de ajuste** para determinar si la muestra de las 200 preferencias de los clientes coincide con la hipótesis nula. La prueba de bondad de ajuste se basa en

la comparación de los resultados muestrales *observados* con los resultados *esperados,* bajo la suposición de que la hipótesis nula es verdadera. Por tanto, el paso siguiente es calcular las preferencias esperadas en los 200 clientes, con el supuesto de que $p_A = 0.30$, $p_B = 0.50$ y $p_C = 0.20$ hacerlo dará los resultados esperados.

Frecuencia observada		
Producto de la empresa A	**Producto de la empresa B**	**Producto de la empresa C**
200(0.30) = 60	200(0.50) = 100	200(0.20) = 40

Como se observa, la frecuencia esperada de cada categoría se encuentra multiplicando el tamaño de la muestra, 200, por la proporción hipotética de esa categoría.

En la prueba de bondad de ajuste lo que interesa son las diferencias entre frecuencias observadas y frecuencias esperadas. Grandes diferencias entre frecuencias observadas y frecuencias esperadas harán dudar sobre la exactitud de las proporciones o participaciones en el mercado hipotéticas. El que las diferencias entre frecuencias observadas y esperadas sean "grandes" o "pequeñas" es una cuestión que se determina con ayuda del estadístico de prueba.

ESTADÍSTICO DE PRUEBA PARA LA PRUEBA DE BONDAD DE AJUSTE

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \tag{12.1}$$

donde

f_i = frecuencia observada en la categoría i

e_i = frecuencia esperada en la categoría i

k = número de categorías

Nota: El estadístico de prueba tiene una distribución chi-cuadrada con $k - 1$ grados de libertad, siempre que en todas las categorías las frecuencias esperadas sean *5 o más*.

Ahora, de regreso con Scott Marketing Research, los datos muestrales se emplearán para probar la hipótesis de que en la población multinomial las proporciones sigan siendo $p_A = 0.30$, $p_B = 0.50$ y $p_C = 0.20$. El nivel de significancia que se va a usar es 0.05. Mediante las frecuencias observadas y esperadas se calcula el valor del estadístico de prueba. Como las frecuencias esperadas son todas 5 o más, se calcula el estadístico de prueba chi-cuadrada como se muestra en la tabla 12.1. Se obtiene $\chi^2 = 7.34$.

La prueba de bondad de ajuste siempre es una prueba de una cola, en la que el rechazo se presenta en la cola superior de la distribución chi-cuadrada.

La hipótesis nula se rechaza si las diferencias entre las frecuencias observadas y esperadas son *grandes*. Diferencias grandes entre las frecuencias esperadas y observadas darán un valor grande del estadístico de prueba. Entonces, la prueba de bondad de ajuste siempre será una prueba de la cola superior. El área en la cola superior se emplea en el método del estadístico de prueba y en el método del valor-*p* para determinar si se puede rechazar la hipótesis nula. Para $k - 1 = 3 - 1 = 2$ grados de libertad, en la tabla de la distribución chi-cuadrada se observan los datos siguientes:

En la sección 11.1 se presenta una introducción a la distribución chi-cuadrada y al uso de la tabla de la distribución chi-cuadrada.

Área en la cola superior	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
Valor χ^2 (2 *gl*)	4.605	5.991	7.378	9.210	10.597

$\chi^2 = 7.34$ (entre 5.991 y 7.378)

TABLA 12.1 CÁLCULO DEL ESTADÍSTICO DE PRUEBA CHI-CUADRADA PARA EL ESTUDIO DE PARTICIPACIÓN DE MERCADO REALIZADO POR SCOTT MARKETING RESEARCH

Categoría	Proporción hipotética	Frecuencia observada (f_i)	Frecuencia esperada (e_i)	Diferencia $(f_i - e_i)$	Cuadrado de la diferencia $(f_i - e_i)^2$	Cuadrado de la diferencia dividido entre frecuencia esperada $(f_i - e_i)^2/e_i$
Empresa A	0.30	48	60	−12	144	2.40
Empresa B	0.50	98	100	−2	4	0.04
Empresa C	0.20	54	40	14	196	4.90
Total		200				$\chi^2 = 7.34$

El estadístico de prueba $\chi^2 = 7.34$ se encuentra entre 5.991 y 7.378. Por consiguiente, el área correspondiente en la cola superior o valor-p debe estar entre 0.05 y 0.025. Como el valor-$p \leq \alpha = 0.05$, se rechaza H_0 y se concluye que la introducción del nuevo producto de la empresa C sí modifica la estructura de la participación de mercado. Con los procedimientos de Excel y Minitab, que se presentan en el apéndice F, al final del libro, se obtiene que si $\chi^2 = 7.34$, el valor-p = 0.0255.

En lugar del método del valor-p se puede utilizar el método del valor crítico con el que se llega a la misma conclusión. Como $\alpha = 0.05$ y los grados de libertad son 2, el valor crítico para el estadístico de prueba es $\chi^2_{0.05} = 5.991$. La regla de rechazo de la cola superior es

$$\text{Rechazar } H_0 \text{ si } \chi^2 \geq 5.991$$

Como $7.34 > 5.991$, se rechaza H_0. Con el método del valor crítico o con el método del valor-p se llega a la misma conclusión.

Aunque no se obtienen más conclusiones como resultado de la prueba, es posible comparar las frecuencias observadas y las frecuencias esperadas de manera informal para tener una idea de cómo ha cambiado la estructura de la participación en el mercado. Se observa que para la empresa C, la frecuencia observada, que es 54, es mayor que la frecuencia esperada, 40. Como la frecuencia esperada estaba basada en la participación existente en el mercado, que la frecuencia observada sea mayor indica que el nuevo producto de la empresa C tendrá un efecto positivo sobre la participación en el mercado de esta empresa. Comparando las frecuencias observadas y esperadas de las otras dos empresas, se observa que la empresa C gana en participación en el mercado afectando más a la empresa A que a la empresa B.

A continuación se presentan, en forma resumida, los pasos que se siguen para realizar una prueba de bondad de ajuste para una distribución poblacional multinomial hipotética.

DISTRIBUCIÓN MULTINOMIAL DE PRUEBAS DE BANDA DE AJUSTE: RESUMEN

1. Establecer las hipótesis nula y alternativa:

   H_0: La población tiene una distribución multinomial con la probabilidad especificada de cada una de las k categorías

   H_a: La población tiene una distribución multinomial con la probabilidad no especificada de cada una de las k categorías

2. Seleccionar una muestra aleatoria y anotar la frecuencia observada f_i en cada categoría.
3. Suponer que la hipótesis nula es verdadera y determinar la frecuencia esperada e_i en cada categoría multiplicando la probabilidad de esa categoría por el tamaño de la muestra.

4. Calcular el valor del estadístico de prueba.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

5. Regla de rechazo:

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $\chi^2 \geq \chi^2_\alpha$

donde α es el nivel de significancia utilizado para la prueba y se tienen $k - 1$ grados de libertad.

Ejercicios

Métodos



1. Probar la hipótesis siguiente usando la prueba de bondad de ajuste χ^2.

$$H_0: p_A = 0.40, p_B = 0.40, \text{ y } p_C = 0.20$$
$$H_a: \text{Las proporciones poblacionales no son } p_A = 0.40, p_B = 0.40, \text{ y } p_C = 0.20$$

En una muestra de 200 elementos se tiene que 60 pertenecen a la categoría A, 120 a la categoría B y 20 a la categoría C.
Use $\alpha = 0.01$ y pruebe si las proporciones son las afirmadas en H_0.
a. Use el método del valor-p.
b. Repita la prueba usando el método del valor crítico.

2. Suponga que tiene una población multinomial con cuatro categorías: A, B, C y D. La hipótesis nula es que la proporción de elementos es la misma en todas las categorías. La hipótesis nula es

$$H_0: p_A = p_B = p_C = p_D = 0.25$$

En la muestra, que es de 300, se obtienen los resultados siguientes:

A: 85 B: 95 C: 50 D: 70

Use $\alpha = 0.05$ para determinar si se rechaza H_0. ¿Cuál es el valor-p?

Aplicaciones



3. Durante las primeras 13 semanas, se registraron las proporciones siguientes de televidentes los sábados de 8 a 9 de la noche: ABC 29%, CBS 28%, NBC 25% e independientes 18%. Dos semanas después en una muestra de 300 hogares se obtuvieron las audiencias siguientes en sábado por la noche: ABC 95 hogares, CBS 70 hogares, NBC 89 hogares e independientes 46 hogares. Use $\alpha = 0.05$ para determinar si han variado las proporciones en la audiencia de televidentes.

4. M&M/Mars, fabricantes de los chocolates M&M, realizaron un sondeo nacional en el que más de 10 millones de personas dieron su preferencia para un nuevo color. El resultado de este sondeo fue el reemplazo de un color café claro por uno azul. En el prospecto “Colors” de M&M/Mars, la distribución de los colores de estos chocolates es la siguiente:

Café	Amarillo	Rojo	Anaranjado	Verde	Azul
30%	20%	20%	10%	10%	10%

En un estudio posterior se emplearon como muestras bolsas de 1 libra para determinar si los porcentajes dados eran reales. En la muestra de 506 dulces los resultados encontrados fueron los siguientes.

Café	Amarillo	Rojo	Anaranjado	Verde	Azul
177	135	79	41	36	38

Use $\alpha = 0.05$ para determinar si estos datos coinciden con los datos dados por la empresa.

5. ¿Dónde es más frecuente que las mujeres compren ropa informal? Según la base de datos de U.S. Shopper se obtuvieron los porcentajes siguientes acerca de las compras de ropa que realizan las mujeres en cada uno de los distintos tipos de tiendas.

Tienda	Porcentaje	Tienda	Porcentaje
Wal-Mart	24	Kohl's	8
Tiendas departamentales tradicionales	11	Por correo	12
JC Penney	8	Otras	37

La categoría otras comprende tiendas como Target, Kmart y Sears, así como numerosas tiendas especializadas. Ninguna de las tiendas de este grupo tiene más de 5% de las compras femeninas. En Atlanta, Georgia, un estudio reciente en el que se usó una muestra de 140 mujeres, los datos encontrados fueron Wal-Mart, 42; Tiendas departamentales tradicionales, 20; JC Penny, 8; Kohl´s, 10; por correo, 21; otras, 39. ¿Estos datos muestran que en Atlanta las compras femeninas difieren de las preferencias que indica la base de datos de U.S. Shopper? ¿Cuál es el valor-*p*? ¿Cuál es la conclusión?

6. La American Bankers Association recoge datos sobre el uso de tarjetas de crédito, tarjetas de débito, efectivo y cheques personales en el pago de compras en tienda (*The Wall Street Journal*, 16 de diciembre de 2003). En 1999, los datos encontrados fueron los siguientes:

Compras en tienda	Porcentaje
Tarjeta de crédito	22
Tarjeta de débito	21
Cheque personal	18
Efectivo	39

En una muestra tomada en el 2003, en 220 compras en tienda se encontró que en 46 se usó tarjeta de crédito, en 67 se usó tarjeta de débito, en 33 se usó un cheque personal y en 74 se pagó en efectivo.

a. Con $\alpha = 0.01$ ¿puede concluir que en este periodo de cuatro años, de 1999 a 2003, ha habido un cambio en la manera en que los clientes pagan sus compras en las tiendas? ¿Cuál es el valor-*p*?
b. A partir de los datos muestrales del 2003, calcule el porcentaje de uso de cada método de pago. ¿Cuál parece haber sido el principal o los principales cambios ocurridos en este periodo de cuatro años?
c. En 2003, ¿qué porcentaje de los pagos se hicieron con tarjeta (tarjeta de crédito o débito)?

7. En el cuadro de accionistas de *The Wall Street Journal* se sigue el comportamiento de las 1 000 empresas principales de Estados Unidos (*The Wall Street Journal*, 10 de marzo de 2003). El comportamiento de cada empresa se califica con base en los rendimientos anuales totales, que comprenden cambios en los precios de las acciones y reinversión de dividendos. Las calificaciones se asignan dividiendo las 1000 empresas en 5 grupos, del A (20% mejor), B (siguiente 20%), hasta E (20% inferior). Lo que se muestra a continuación son las calificaciones en un año obtenidas por las 60 empresas más grandes. ¿El comportamiento de las empresas más grandes difiere de las 1000 empresas del cuadro de accionistas?

A	B	C	D	E
5	8	15	20	12

8. ¿Qué tan bueno es el servicio que dan las líneas aéreas a sus clientes? En un estudio las evaluaciones dadas por los clientes fueron las siguientes: 3% excelente, 28% bueno, 45% aceptable y 24% malo (*BusinessWeek*, 11 de septiembre de 2000). En otro estudio sobre las empresas de servicio telefónico, en una muestra de 400 adultos las evaluaciones fueron las siguientes: 24 excelente, 124 bueno, 172 aceptable y 80 malo. ¿La distribución de las evaluaciones a las empresas telefónicas difiere de la distribución de las evaluaciones a las líneas aéreas? Emplee $\alpha = 0.01$. ¿Cuál es su conclusión?



Prueba de independencia

Otra aplicación importante de la distribución chi-cuadrada es el empleo de datos muestrales para probar la independencia de dos variables. Para ilustrar la prueba de independencia se considerará la prueba de independencia realizada por la Alber's Brewery de Tucson, Arizona. Alber's produce y distribuye tres tipos de cerveza: ligera, clara y oscura. Al analizar los segmentos de mercado de las tres cervezas, el grupo de investigación de mercado de la empresa se preguntó si las preferencias de los consumidores por estos tipos de cerveza diferían entre hombres y mujeres. En caso de que las preferencias fueran independientes del género del consumidor, iniciarían una campaña publicitaria para todas las cervezas de Alber's. Pero, si las preferencias por los distintos tipos de cerveza dependían del género del consumidor, la empresa ajustaría sus promociones a los mercados.

Para determinar si la preferencia por un tipo de cerveza (ligera, clara u oscura) era independiente del género del consumidor (hombre o mujer) se usó una prueba de independencia. Las hipótesis para esta prueba de independencia fueron:

H_0: La preferencia por un tipo de cerveza es independiente del género del consumidor.
H_a: La preferencia por un tipo de cerveza no es independiente del género del consumidor.

Para describir la situación a estudio se usa la tabla 12.2. Después de identificar la población como todos los consumidores de cerveza, hombres y mujeres, se toma una muestra y a cada individuo

TABLA 12.2 TABLA DE CONTINGENCIA DE CERVEZA PREFERIDA Y GÉNERO DEL CONSUMIDOR

		Cerveza preferida		
		Ligera	**Clara**	**Oscura**
Género	**Hombre**	celda (1,1)	celda (1,2)	celda (1,3)
	Mujer	celda (2,1)	celda (2,2)	celda (2,3)

TABLA 12.3 RESULTADOS MUESTRALES DEL TIPO DE CERVEZA QUE PREFIEREN HOMBRES Y MUJERES (FRECUENCIAS OBSERVADAS)

		Cerveza preferida			
		Ligera	**Clara**	**Oscura**	**Total**
Género	**Hombre**	20	40	20	80
	Mujer	30	30	10	70
	Total	50	70	30	150

Para probar si dos variables son independientes, se toma una muestra y se usa una tabulación cruzada para resumir los datos de las dos variables simultáneamente.

se le pide que indique cuál de las tres cervezas de Alber's prefiere. Cada individuo de la muestra pertenecerá a una de las seis celdas de la tabla. Así, por ejemplo, se puede tener un individuo que sea hombre y que prefiera la cerveza clara (celda (1,2)), o una mujer que prefiera la cerveza ligera (celda (2,1)), o una mujer que prefiera la cerveza oscura (celda (2,3)), etc. Dado que en la tabla se han enumerado todas las posibles combinaciones de cerveza preferida y género o, en otras palabras, todas las posibles contingencias, a la tabla 12.2 se le llama **tabla de contingencia**. Como en la prueba de independencia se usa el formato de las tablas de contingencia, a esta prueba también se le suele llamar *prueba de tabla de contingencia*.

Suponga que toma una muestra aleatoria simple de 150 consumidores de cerveza. Cada individuo de la muestra prueba los tres tipos de cerveza y después se le pide que indique cuál prefiere o cuál es su primera elección. En la tabulación cruzada de la tabla 12.3 se presentan las respuestas obtenidas en el estudio. Como se ve, los datos para la prueba de independencia se obtienen contando las cantidades o frecuencias correspondientes a cada celda o categoría. De las 150 personas que formaban la muestra, 20 hombres prefirieron la cerveza ligera, 40 hombres prefirieron la cerveza clara, 20 hombres prefirieron la cerveza oscura, etcétera.

Los datos de la tabla 12.3 son las frecuencias observadas para cada una de las seis clases o categorías. Si determina las frecuencias esperadas bajo la suposición de independencia entre cerveza preferida y género del consumidor, se puede emplear la distribución chi-cuadrada para establecer si existe diferencia significativa entre las frecuencias observadas y las esperadas.

Las frecuencias esperadas para las celdas de la tabla de contingencia se basan en la idea siguiente. Primero se supone que la hipótesis nula es verdadera, es decir, que la cerveza preferida es independiente del género del consumidor. Después se observa que en toda la muestra de 150 consumidores de cerveza, 50 prefirieron la cerveza ligera, 70 prefirieron la cerveza clara, y 30 prefirieron la cerveza oscura. En términos de proporciones se concluye que $^{50}/_{150} = ^{1}/_{3}$ de los consumidores prefirió la cerveza ligera, $^{70}/_{150} = ^{7}/_{15}$ prefirieron la cerveza clara y $^{30}/_{150} = ^{1}/_{5}$ prefirió la cerveza oscura. Si la suposición de *independencia* es correcta, estas proporciones serán las que se observen tanto entre los hombres como entre las mujeres. Por consiguiente, bajo la suposición de independencia, es de esperarse que en la muestra de 80 consumidores del sexo masculino, $(^{1}/_{3})80 = 26.67$ prefieran la cerveza ligera, $(^{7}/_{15})80 = 37.33$ prefieran la cerveza clara y $(^{1}/_{5})80 = 16$ prefieran la cerveza oscura. Aplicando las proporciones correspondientes a los 70 consumidores del sexo femenino, se obtienen las frecuencias esperadas que se muestran en la tabla 12.4.

TABLA 12.4 FRECUENCIAS ESPERADAS SI LA PREFERENCIA POR UNO DE LOS TIPOS DE CERVEZA ES INDEPENDIENTE DEL GÉNERO DEL CONSUMIDOR

		Cerveza preferida			
		Ligera	**Clara**	**Oscura**	**Total**
Género	**Hombre**	26.67	37.33	16.00	80
	Mujer	23.33	32.67	14.00	70
	Total	50.00	70.00	30.00	150

Sea e_{ij} la frecuencia esperada en el renglón i columna j de la tabla de contingencia. Mediante dicha notación, ahora se reconsidera el cálculo de la frecuencia esperada correspondiente a los hombres (renglón $i = 1$) que prefieren la cerveza clara (columna $j = 2$); es decir, la frecuencia esperada e_{12}. Siguiendo el argumento anterior para el cálculo de las frecuencias esperadas, se ve que

$$e_{12} = (7/15)80 = 37.33$$

Expresión que se formula de una manera ligeramente diferente como

$$e_{12} = (7/15)80 = (70/150)80 = \frac{(80)(70)}{150} = 37.33$$

Observe que en esta expresión, 80 es el número total de hombres (total del renglón 1), 70 es la cantidad total de individuos que prefieren la cerveza clara (total de la columna 2) y 150 es el tamaño total de la muestra. De lo que se ve que

$$e_{12} = \frac{(\text{Total del renglón 1})\ (\text{Total de la columna 2})}{\text{Tamaño de la muestra}}$$

La generalización de esta expresión lleva a la fórmula siguiente para obtener las frecuencias esperadas en una tabla de contingencia para una prueba de independencia.

FRECUENCIAS ESPERADAS EN UNA TABLA DE CONTINGENCIA BAJO LA SUPOSICIÓN DE INDEPENDENCIA

$$e_{ij} = \frac{(\text{Total del renglón } i)\ (\text{Total de la columna } j)}{\text{Tamaño de la muestra}} \qquad \textbf{(12.2)}$$

Al aplicar esta fórmula para los consumidores hombres que prefieren cerveza oscura, se encuentra que la frecuencia esperada es $e_{13} = (80)(30)/150 = 16.00$, como se muestra en la tabla 12.4. Use la ecuación 12.2 para verificar las otras frecuencias esperadas que se presentan en la tabla 12.4.

El procedimiento de prueba para comparar las frecuencias esperadas de la tabla 12.4 con las frecuencias observadas de la tabla 12.3 es semejante a los cálculos para la prueba de bondad de ajuste de la sección 12.1. En concreto, el valor χ^2 que se basa en frecuencias observadas y esperadas se calcula como se indica a continuación.

ESTADÍSTICO DE PRUEBA PARA INDEPENDENCIA

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \qquad \textbf{(12.3)}$$

donde

f_{ij} = frecuencia observada en la categoría del renglón i columna j de la tabla de contingencia

e_{ij} = frecuencia esperada en la categoría del renglón i columna j de la tabla de contingencia, basada en la suposición de independencia.

Nota: Si una tabla de contingencia tiene n renglones y m columnas, el estadístico de prueba tiene una distribución chi-cuadrada con $(n - 1)(m - 1)$ grados de libertad, siempre y cuando en todas las categorías las frecuencias esperadas sean cinco o más.

TABLA 12.5 CÁLCULO DEL ESTADÍSTICO DE PRUEBA CHI-CUADRADA PARA DETERMINAR SI LA PREFERENCIA POR UN TIPO DE CERVEZA ES INDEPENDIENTE DEL GÉNERO DEL CONSUMIDOR

Género	Cerveza preferida	Frecuencia observada (f_{ij})	Frecuencia esperada (e_{ij})	Diferencia ($f_{ij} - e_{ij}$)	Cuadrado de la diferencia ($f_{ij} - e_{ij})^2$	Cuadrado de la diferencia dividido entre frecuencia esperada $(f_{ij} - e_{ij})^2/e_{ij}$
Hombre	Ligera	20	26.67	−6.67	44.44	1.67
Hombre	Clara	40	37.33	2.67	7.11	0.19
Hombre	Oscura	20	16.00	4.00	16.00	1.00
Mujer	Ligera	30	23.33	6.67	44.44	1.90
Mujer	Clara	30	32.67	−2.67	7.11	0.22
Mujer	Oscura	10	14.00	−4.00	16.00	1.14
	Total	150				$\chi^2 = 6.12$

La doble sumatoria que aparece en la ecuación (12.3) indica que el cálculo debe hacerse con todas las celdas que aparecen en la tabla de contingencia.

En las frecuencias esperadas que aparecen en la tabla 12.4, se ve que en cada categoría la frecuencia esperada es de 5 o más. Por tanto se puede proceder a calcular el estadístico de prueba chi-cuadrada. En la tabla 12.5 se presentan los cálculos necesarios para obtener el estadístico de prueba chi-cuadrada que se utiliza para determinar si la preferencia por una cerveza es independiente del género del consumidor. Como se observa, el valor del estadístico de prueba es $\chi^2 = 6.12$.

La prueba de independencia siempre es una prueba de una cola, en la que la región de rechazo se encuentra en la cola superior de la distribución chi-cuadrada.

El número de grados de libertad para la distribución chi-cuadrada adecuada se obtiene multiplicando el número de renglones menos 1 por el número de columnas menos 1. Como se tienen dos renglones y tres columnas, los grados de libertad son $(2 - 1)(3 - 1) = 2$. Como ocurre en la prueba de bondad de ajuste, en la prueba de independencia se rechaza H_0 si las diferencias entre frecuencias observadas y esperadas dan un valor grande del estadístico de prueba. De manera que la prueba de independencia es también una prueba de la cola superior. La tabla de la distribución chi-cuadrada (tabla 3 del apéndice B), proporciona la información siguiente para 2 grados de libertad.

Área en la cola superior	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
Valor χ^2 (2 *df*)	4.605	5.991	7.378	9.210	10.597

$\chi^2 = 6.12$

El estadístico de prueba, $\chi^2 = 6.12$, se encuentra entre 5.991 y 7.378. Por tanto, el área correspondiente en la cola superior o valor-p está entre 0.05 y 0.025. Empleando los procedimientos de Minitab o de Excel que se presentan en el apéndice F, se obtiene que, valor-p = 0.0469. Como el valor-$p \leq \alpha$ 0.05, se rechaza la hipótesis nula y se concluye que la preferencia por una cerveza no es independiente del género del consumidor.

Para simplificar los cálculos que se requieren en una prueba de independencia se usan paquetes de software como Minitab o Excel. La información a suministrar en estos procedimientos es la tabla de contingencia con las frecuencias observadas como se muestran en la tabla 12.3. El software calcula automáticamente las frecuencias esperadas, el valor del estadístico de prueba χ^2 y el valor-p. En los apéndices 12.1 y 12.2 se presentan los procedimientos de Minitab y de Excel para esta prueba de independencia. En la figura 12.1 aparecen los resultados que da Minitab para la prueba de la Alber's Brewery.

Mediante una comparación informal de las frecuencias observadas y esperadas se obtiene una idea de la dependencia entre cerveza preferida y género. Al observar las tablas 12.3 y 12.4 resalta que en los consumidores de sexo masculino las frecuencias observadas en la preferencia por cervezas clara y oscura son más altas que las frecuencias esperadas, mientras que en las mu-

FIGURA 12.1 RESULTADOS DE MINITAB PARA LA PRUEBA DE INDEPENDENCIA DE LA ALBER'S BREWERY

```
Expected counts are printed below observed counts

         Light    Regular     Dark    Total
    1       20         40       20       80
         26.67      37.33    16.00

    2       30         30       10       70
         23.33      32.67    14.00

Total       50         70       30      150

DF = 2, P-Value = 0.047
```

jeres la frecuencia observada en la preferencia por cerveza ligera es mayor que la frecuencia esperada. Dichas observaciones permiten comprender las diferentes preferencias por cerveza entre los hombres y las mujeres.

A continuación se resumen los pasos para una prueba de tabla de contingencia para independencia.

PRUEBA DE INDEPENDENCIA: RESUMEN

1. Establecer las hipótesis nula y alternativa.

H_0: La variable de las columnas es independiente de la variable de los renglones
H_a: La variable de las columnas no es independiente de la variable de los renglones

2. Seleccionar una muestra aleatoria y anotar en cada celda de la tabla de contingencias las frecuencias observadas.
3. Emplear la ecuación (12.2) para calcular las frecuencias esperadas de cada celda.
4. Utilizar la ecuación (12.3) para calcular el valor del estadístico de prueba.
5. Regla de rechazo:

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $\chi^2 \geq \chi^2_\alpha$

donde α es el nivel de significancia, y los n renglones y las m columnas dan los $(n-1)(m-1)$ grados de libertad.

NOTAS Y COMENTARIOS

El estadístico de prueba para las pruebas chi-cuadrada de este capítulo requiere una frecuencia esperada de cinco en cada categoría. Si en una categoría la frecuencia esperada es menor que cinco, es conveniente combinar dos categorías adyacentes para tener una frecuencia esperada de cinco o más en cada categoría.

Ejercicios

Métodos



9. La siguiente tabla de contingencia 2 × 3 contiene las frecuencias observadas en una muestra de tamaño 200. Pruebe la independencia de las variables de renglón y de columna usando la prueba χ^2 con $\alpha = 0.05$.

Variable de los renglones	Variable de las columnas A	B	C
P	20	44	50
Q	30	26	30

10. La siguiente tabla de contingencia 3 × 3 contiene las frecuencias observadas en una muestra de 240. Pruebe la independencia de la variable de los renglones y la variable en las columnas usando la prueba χ^2 con $\alpha = 0.05$.

Variable de los renglones	Variable de las columnas A	B	C
P	20	30	20
Q	30	60	25
R	10	15	30

Aplicaciones



11. Una de las preguntas a los suscriptores de *BusinessWeek* fue, “En sus viajes de negocios de los últimos 12 meses, ¿qué tipo de boleto de avión ha comprado?” Los datos obtenidos se presentan en la tabla de contingencia siguiente.

Tipo de boleto	Tipo de vuelo Vuelo nacional	Vuelo internacional
Primera clase	29	22
Clase negocios/ejecutivo	95	121
Vuelo tradicional/clase económica	518	135

Use $\alpha = 0.05$ y pruebe la independencia entre tipo de vuelo y tipo de boleto. ¿Cuál es la conclusión?

12. Visa Card USA estudió la frecuencia con que los consumidores de diversos rangos de edades usan tarjetas plásticas (de crédito o de débito) al pagar sus compras (Associated Press, 16 de enero de 2006). A continuación se presentan los datos muestrales de 300 clientes divididos en cuatro grupos de edades.

Forma de pago	Grupo de edad 18–24	25–34	35–44	45 y más
Plástico	21	27	27	36
Efectivo o cheque	21	36	42	90

a. Pruebe la independencia entre el método de pago y el grupo de edad. ¿Cuál es el valor-*p*? Usando $\alpha = 0.05$, ¿cuál es su conclusión?
b. Si la forma de pago y el grupo de edad no son independientes, ¿qué observación puede hacer acerca de la diferencia en el uso de plástico en los diversos grupos de edades?
c. ¿Qué consecuencias tiene este estudio para empresas como Visa, MasterCard y Discover?

13. Dados los incrementos porcentuales anuales de dos dígitos en los costos de los seguros médicos (en Estados Unidos), cada día más trabajadores carecen de un seguro de esta naturaleza (*USA Today*, 23 de enero de 2004). Los datos muestrales siguientes proporcionan una comparación entre los trabajadores con y sin seguro médico en empresas pequeñas, medianas y grandes. Para los propósitos de este estudio, empresas pequeñas son empresas que tienen menos de 100 emplea-

dos. Empresas medianas son empresas que tienen de 100 a 999 empleados y empresas gran des son empresas que tienen 1000 o más empleados. Los datos muestrales corresponden a 50 empleados de empresas pequeñas, 75 empleados de empresas medianas y 100 empleados de empresas grandes.

	Seguro médico		
Tamaño de la empresa	**Sí**	**No**	**Total**
Pequeño	36	14	50
Mediano	65	10	75
Grande	88	12	100

a. Realice una prueba de independencia para determinar si tener un seguro médico es independiente del tamaño de la empresa.
b. El artículo de *USA Today* considera más probable que los empleados de empresas pequeñas carezcan de un seguro médico. Use porcentajes basados en la tabla anterior para apoyar dicha conclusión.

14. Un estudio del Public Interest Research Group (PIRG) del estado de Washington indica que 46% de los estudiantes universitarios de tiempo completo trabaja 25 o más horas por semana. El estudio del PIRG proporciona datos sobre los efectos del trabajo en las calificaciones *USA Today*, 17 de abril de 2002). En este estudio, de 200 estudiantes que conformaban la muestra, 90 trabajaban 1-15 horas por semana, 60 trabajaban 16-24 horas por semana y 50 trabajaban 25-34 horas por semana. A continuación se presentan las cantidades muestrales de estudiantes que indicaron que su trabajo tenía un efecto positivo, ningún efecto o un efecto negativo sobre sus calificaciones.

	Efecto sobre las calificaciones			
Horas trabajadas por semana	**Positivo**	**Ninguno**	**Negativo**	**Total**
1–15 horas	26	50	14	90
16–24 horas	16	27	17	60
25–34 horas	11	19	20	50

a. Realice una prueba de independencia para determinar si el efecto sobre las calificaciones es independiente de las horas trabajadas por semana. Use $\alpha = 0.05$. ¿Cuál es el valor-p y cuál es su conclusión?
b. Use porcentajes de renglón para conocer más acerca del efecto del trabajo sobre las calificaciones.

15. FlightStats, Inc., recoge datos sobre el número de vuelos programados y el número de vuelos efectuados en los principales aeropuertos de Estados Unidos. Los datos de FlightStats indican que 56% de los vuelos programados en los aeropuertos de Newark, La Guardia y Kennedy se efectuaron durante una tormenta de nieve que duró tres días (*The Wall Street Journal*, 21 de febrero de 2006). Todas las aerolíneas afirman que siempre operan dentro de parámetros de seguridad preestablecidos: si las condiciones son muy malas, no vuelan. Los siguientes datos presentan una muestra de 400 vuelos programados durante tormentas de nieve.

	Aerolínea				
¿Voló?	**American**	**Continental**	**Delta**	**United**	**Total**
Sí	48	69	68	25	210
No	52	41	62	35	190

Use la prueba de independencia chi-cuadrada y 0.05 como nivel de significancia para analizar estos datos. ¿Cuál es la conclusión? ¿Qué aerolínea elegiría para volar en semejantes condiciones de tormentas de nieve? Explique.

16. En los negocios cada vez se hacen más pedidos en línea. Una asociación recabó datos sobre la proporción de órdenes electrónicas llenadas correctamente de acuerdo con el tipo de industria (*Investor's Business Daily*, 8 de mayo de 2000). En una muestra de 700 órdenes electrónicas se obtuvieron los resultados siguientes.

	Industria			
Orden	**Farmacéutica**	**De consumo**	**Computadoras**	**Telecomunicación**
Correcta	207	136	151	178
Incorrecta	3	4	9	12

a. Haga una prueba de hipótesis para determinar si el llenado correcto de las órdenes es independiente de la industria. Use $\alpha = 0.05$. ¿Cuál es la conclusión?
b. ¿Qué industria tiene el porcentaje más alto de órdenes llenadas correctamente?

17. La National Sleep Foundation realiza encuestas para determinar si las horas de sueño por noche son independientes de la edad (*Newsweek*, 19 de enero de 2004). Las siguientes son las horas de sueño entre semana en una muestra de personas de 49 años o menos y en otra muestra de personas de 50 años o más.

	Horas de sueño				
Edad	**Menos de 6**	**6 a 6.9**	**7 a 7.9**	**8 o más**	**Total**
49 o menos	38	60	77	65	240
50 o más	36	57	75	92	260

a. Realice una prueba de independencia para determinar si las horas de sueño entre semana son independientes de la edad. Use $\alpha = 0.05$. ¿Cuál es el valor-p y cuál es la conclusión?
b. Dé una estimación del porcentaje de personas que duermen menos de 6 horas, de 6 a 6.9 horas, de 7 a 7.9 horas y 8 horas o más.

18. Muestras tomadas en tres ciudades, Anchorage, Atlanta y Miniápolis, se usaron para obtener información acerca del porcentaje de parejas casadas en las que los dos cónyuges trabajan (*USA Today*, 15 de enero de 2006). Analice los datos siguientes para ver si el hecho de que los dos cónyuges trabajen es independiente del lugar donde viven. Use 0.05 como nivel de significancia. ¿Cuál es su conclusión? Dé la estimación general del porcentaje de parejas casadas en las que ambos cónyuges trabajan.

	Ubicación		
Trabajan	**Anchorage**	**Atlanta**	**Minneapolis**
Ambos	57	70	63
Sólo uno	33	50	90

19. En un programa de televisión los dos presentadores suelen dar la impresión de no estar en absoluto de acuerdo al evaluar películas. En la evaluación de una película pueden estar a favor ("pul-

gar hacia arriba”), en contra (“pulgar hacia abajo”) o indiferente. Se presentan las evaluaciones de 160 películas hechas por los dos presentadores.

	Presentador B		
Presentador A	**A favor**	**Indiferente**	**En contra**
A favor	24	8	13
Indiferente	8	13	11
En contra	10	9	64

Para analizar estos datos use la prueba chi-cuadrada de independencia con 0.01 como nivel de significancia. ¿Cuál es la conclusión?



Prueba de bondad de ajuste: distribuciones de Poisson y normal

En la sección 12.1 se introdujo la prueba de bondad de ajuste para poblaciones multinomiales. En general, la prueba de bondad de ajuste puede usarse con cualquier distribución de probabilidad hipotética. En esta sección se ilustra el uso de la prueba de bondad de ajuste para el caso en que se tiene la hipótesis de que la población tiene una distribución de Poisson o una distribución normal. Como verá, en la prueba de bondad de ajuste y en el uso de la distribución chi-cuadrada se sigue el mismo procedimiento general aplicado para la prueba de bondad de ajuste de la sección 12.1.

Distribución de Poisson

El uso de la prueba de bondad de ajuste se ilustra en el caso de una distribución poblacional que hipotéticamente tiene una distribución de Poisson. Considere, por ejemplo, las llegadas de los clientes al Dubek’s Food Market en Tallase, Florida. Dado que recién ha habido algunos problemas de personal, los gerentes solicitan los servicios de una empresa de consultoría para que les ayude en la programación de los empleados de cajas. Después de revisar el avance de las filas en las cajas, la empresa de consultoría sugerirá un procedimiento para la programación de los empleados de cajas. Este procedimiento se basa en un análisis matemático de las filas y sólo es aplicable si el número de llegadas de clientes durante un determinado lapso de tiempo sigue una distribución de Poisson. Por tanto, antes de poner en marcha el procedimiento de programación, habrá que recolectar datos sobre las llegadas de los clientes y realizar una prueba estadística para ver si es razonable suponer que las llegadas de los clientes siguen una distribución de Poisson.

Las llegadas de los clientes a la tienda se definen en términos de *cantidad de clientes* que entran en la tienda durante intervalos de 5 minutos. Por tanto las hipótesis nula y alternativa en este estudio son las siguientes:

H_0: La cantidad de clientes que entran en la tienda durante intervalos de 5 minutos tiene una distribución de probabilidad de Poisson

H_a: La cantidad de clientes que entran en la tienda durante intervalos de 5 minutos no tienen una distribución de probabilidad de Poisson

Si una muestra de llegadas de clientes indica que no se puede rechazar H_0, Dubeck’s procederá a poner en marcha el proceso de programación de la empresa de consultoría. Pero, si la muestra lleva a rechazar H_0, no se podrá suponer que las llegadas siguen una distribución de Poisson y habrá que considerar otro procedimiento de programación.

Para probar la suposición de que las llegadas de los clientes en las mañanas de los días entre semana siguen una distribución de Poisson, un empleado de la tienda toma una muestra aleatoria de 128 intervalos de 5 minutos, en las mañanas de tres semanas consecutivas. Durante cada uno de los intervalos de 5 minutos que forman la muestra, el empleado registra el número de lle-

TABLA 12.6

Frecuencias observadas en las llegadas de los clientes de Dubek's en una muestra de 128 intervalos de 5 minutos

Número de llegadas de clientes	Frecuencia observada
0	2
1	8
2	10
3	12
4	18
5	22
6	22
7	16
8	12
9	6
Total	128

gadas de clientes. Para resumir los datos, el empleado determina el número de intervalos de 5 minutos en los que no hubo ninguna llegada, el número de intervalos de 5 minutos en los que hubo una llegada, el número de intervalos de 5 minutos en los que hubo dos llegadas, etc. Estos datos se presentan en la tabla 12.6

La tabla 12.6 da las frecuencias observadas en las 10 categorías. Ahora se usa la prueba de bondad de ajuste para determinar si la muestra de los 128 lapsos de tiempo favorecen la hipótesis de que las llegadas tienen una distribución de Poisson. Para usar la prueba de bondad de ajuste, se necesitan considerar, las frecuencias esperadas para cada una de las 10 categorías, bajo la suposición de que la distribución de las llegadas siga una distribución de Poisson. Es decir, si en realidad las llegadas de los clientes siguen una distribución de Poisson, se necesita calcular el número esperado de lapsos de tiempo en los que llegarán cero clientes, un cliente, dos clientes, etcétera.

La función de probabilidad de Poisson, que ya se presentó en el capítulo 5, es

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \tag{12.4}$$

En esta función, μ representa la media o número esperado de llegadas de clientes en lapsos de 5 minutos, x representa la variable aleatoria del número de llegadas de clientes en un lapso de 5 minutos y $f(x)$ es la probabilidad de x llegadas de clientes en un lapso de 5 minutos.

Antes de usar la ecuación (12.4) para calcular las probabilidades de Poisson se necesita una estimación de μ, el número medio de llegadas de clientes en un lapso de 5 minutos. La media muestral de los datos de la tabla 12.6 proporciona dicha estimación. Como se tienen dos lapsos de 5 minutos en los que no llegó ningún cliente, ocho lapsos de 5 minutos en los que llegó un cliente, etc., el número total de llegadas de clientes en los 128 lapsos de 5 minutos es $0(2) + 1(8) + 2(10) + \cdots + 9(6) = 640$. Las 640 llegadas de clientes en los 128 lapsos de tiempo de la muestra dan una media de llegadas $\mu = 640/128 = 5$ llegadas de clientes por lapso de 5 minutos. Con este valor como media para la distribución de Poisson, una estimación de la función de probabilidad de Poisson en el caso de Dubek's es

$$f(x) = \frac{5^x e^{-5}}{x!} \tag{12.5}$$

Esta función de probabilidad puede evaluarse para distintos valores de x y determinar así la probabilidad que corresponde a las diferentes categorías de llegadas. En la tabla 12.7 se presentan tales probabilidades, las cuales pueden encontrarse también en la tabla 7 del apéndice B. Por

TABLA 12.7 FRECUENCIAS ESPERADAS EN LAS LLEGADAS DE LOS CLIENTES A DUBEK'S SUPONIENDO QUE SIGAN UNA DISTRIBUCIÓN DE POISSON CON $\mu = 5$

Número de llegadas de clientes (x)	Probabilidad de Poisson $f(x)$	Número esperado de lapsos de 5 minutos con x llegadas, $128 f(x)$
0	0.0067	0.86
1	0.0337	4.31
2	0.0842	10.78
3	0.1404	17.97
4	0.1755	22.46
5	0.1755	22.46
6	0.1462	18.71
7	0.1044	13.36
8	0.0653	8.36
9	0.0363	4.65
10 o más	0.0318	4.07
	Total	128.00

ejemplo, la probabilidad de que lleguen cero clientes en un lapso de cinco minutos es $f(0) = 0.0067$, la probabilidad de que llegue un cliente en un lapso de 5 minutos es $f(1) = 0.0337$, etc. Como se vio en la sección 12.1, la frecuencia esperada en cada una de las categorías se encuentra multiplicando su probabilidad por el tamaño de la muestra. Por ejemplo, el número de lapsos de tiempo con cero llegadas es $(0.0067)(128) = 0.86$, el número esperado de lapsos de tiempo con una llegada es $(0.0337)(128) = 4.31$, etcétera.

Cuando en alguna categoría el número esperado es menor que cinco, no se satisfacen las condiciones para la prueba χ^2. Cuando esto ocurre, se pueden combinar categorías adyacentes para que el número esperado sea cinco o más.

Antes de hacer los cálculos habituales para comparar las frecuencias observadas y esperadas, hay que observar que en la tabla 12.7, hay cuatro categorías que tienen una frecuencia esperada menor que cinco. Esto viola los requerimientos para el uso de la distribución chi-cuadrada. Sin embargo, no es una dificultad, ya que se pueden combinar categorías menores que cinco para satisfacer la condición de que la frecuencia esperada sea "por lo menos cinco". Aquí, se combinan 0 y 1 en una sola categoría y también se combinan 9 y "10 o más" en una sola categoría. De esta manera se satisface la regla de un mínimo de cinco como frecuencia esperada en cada categoría. En la tabla 12.8 se presentan las frecuencias observadas y las esperadas después de combinar estas categorías.

Como en la sección 12.1, la prueba de bondad de ajuste se centra en las diferencias entre frecuencias observadas y esperadas, $f_i - e_i$. Por tanto, para calcular el estadístico de prueba chi-cuadrada se usarán las frecuencias observadas y las esperadas de la tabla 12.8.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

En la tabla 12.9 se muestran los cálculos necesarios para obtener el valor del estadístico de prueba chi-cuadrada. El valor del estadístico de prueba es $\chi^2 = 10.96$.

En general, en una prueba de bondad de ajuste la distribución chi-cuadrada tiene $k - p - 1$ grados de libertad, donde k es el número de categorías y p es el número de parámetros poblacionales estimados a partir de los datos muestrales. En este caso, como se ve en la tabla 12.9, $k = 9$ categorías. Como los datos muestrales se usaron para estimar la media de la distribución de Poisson, $p = 1$. Por ende, se tiene $k - p - = k - 2$ grados de libertad. Como $k = 9$, se tienen $9 - 2 = 7$ grados de libertad.

Suponga que en la prueba de la hipótesis nula de que la distribución de probabilidad de las llegadas de los clientes es una distribución de Poisson se usa 0.05 como nivel de significancia. Para probar esta hipótesis, se necesita determinar el valor-p correspondiente al valor del estadístico de prueba $\chi^2 = 10.96$ hallando el área en la cola superior de la distribución chi-cuadrada con 7 grados de libertad. En la tabla 3 del apéndice B se encuentra que $\chi^2 = 10.96$ corresponde a un área, en la cola superior, mayor que 0.10. Por tanto se sabe que el valor-p es mayor que 0.10. Con los procedimientos de Minitab y de Excel que se describen en el apéndice F se obtiene que va-

TABLA 12.8 FRECUENCIAS OBSERVADAS Y ESPERADAS EN LAS LLEGADAS DE LOS CLIENTES A DUBEK´S, DESPUÉS DE COMBINAR CATEGORÍAS

Número de llegadas de clientes (x)	Frecuencia observada (f_i)	Frecuencia esperada (e_i)
0 or 1	10	5.17
2	10	10.78
3	12	17.97
4	18	22.46
5	22	22.46
6	22	18.72
7	16	13.37
8	12	8.36
9 o más	6	8.72
Total	128	128.00

TABLA 12.9 CÁLCULO DEL ESTADÍSTICO DE PRUEBA CHI-CUADRADA PARA EL ESTUDIO DE DUBEK'S FOOD MARKET

Número de llegadas de clientes (x)	Frecuencia observada (f_i)	Frecuencia esperada (e_i)	Diferencia ($f_i - e_i$)	Cuadrado de la diferencia $(f_i - e_i)^2$	Cuadrado de la diferencia dividido entre la frecuencia esperada $(f_i - e_i)^2/e_i$
0 o 1	10	5.17	4.83	23.28	4.50
2	10	10.78	−0.78	0.61	0.06
3	12	17.97	−5.97	35.62	1.98
4	18	22.46	−4.46	19.89	0.89
5	22	22.46	−0.46	0.21	0.01
6	22	18.72	3.28	10.78	0.58
7	16	13.37	2.63	6.92	0.52
8	12	8.36	3.64	13.28	1.59
9 o más	6	8.72	−2.72	7.38	0.85
Total	128	128.00			$\chi^2 = 10.96$

lor-p = 0.1404. Como el valor-$p > \alpha = 0.05$, no se puede rechazar H_0. De manera que no se puede rechazar la suposición de que la distribución de probabilidad de las llegadas de los clientes, en las mañanas entre semana, siga una distribución de probabilidad de Poisson. De esta manera, los administradores de Dubek's pueden continuar con el procedimiento de programación para las mañanas de los días entre semana.

PRUEBA DE BONDAD DE AJUSTE PARA LA DISTRIBUCIÓN DE POISSON: RESUMEN

1. Establecer las hipótesis nula y alternativa.

H_0: La población tiene una distribución de Poisson
H_a: La población no tiene una distribución de Poisson

2. Tomar una muestra aleatoria y
   a. Para cada valor de la variable aleatoria de Poisson anotar la frecuencia observada f_i.
   b. Calcular el número medio μ de las ocurrencias.

3. Calcular, para cada valor de la variable aleatoria de Poisson, la frecuencia esperada e_i de ocurrencias. Multiplicar el tamaño de la muestra por la probabilidad de su ocurrencia de cada valor de la variable aleatoria de Poisson. Si para algún valor hay menos de cinco ocurrencias esperadas, combinar valores adyacentes y reducir el número de categorías cuanto sea necesario.

4. Calcular el valor del estadístico de prueba.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

5. Regla de rechazo:

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $\chi^2 \geq \chi^2_\alpha$

donde α es el nivel de significancia y los grados de libertad son $k - 2$.

Distribución normal

La prueba de bondad de ajuste para la distribución normal también se basa en el uso de la distribución chi-cuadrada. Se sigue un procedimiento similar al aplicado para la distribución de Poisson. Las frecuencias observadas en las diversas categorías de los datos muestrales se comparan con las frecuencias esperadas, cuando se supone que la población tiene una distribución normal. Como la distribución normal es continua, es necesario modificar la manera en que se definen las categorías y la forma en que se calculan las frecuencias esperadas. La prueba de bondad de ajuste para una distribución normal se va a ilustrar empleando los datos de los exámenes presentados por las personas que solicitan empleo en la empresa Chemline, Inc.; estos datos se presentan en la tabla 12.10.

TABLA 12.10 PUNTUACIONES OBTENIDAS POR LOS INTEGRANTES DE UNA MUESTRA ALEATORIA DE 50 SOLICITANTES DE EMPLEO EN LA PRUEBA DE APTITUDES DE CHEMLINE

71	66	61	65	54	93
60	86	70	70	73	73
55	63	56	62	76	54
82	79	76	68	53	58
85	80	56	61	61	64
65	62	90	69	76	79
77	54	64	74	65	65
61	56	63	80	56	71
79	84				



Cada año Chemline contrata cerca de 400 empleados nuevos para sus cuatro fábricas en Estados Unidos. El director de personal se pregunta si la población de las puntuaciones en los exámenes de los solicitantes tendrá una distribución normal. Si es así, esta distribución podría servir para evaluar las puntuaciones; es decir, podrían identificarse fácilmente las puntuaciones en 20% superior, en 40% inferior, etc. Por tanto, se desea probar la hipótesis nula de que la población de las puntuaciones de estos exámenes tiene una distribución normal.

Para empezar, se obtendrán estimaciones de la media y de la desviación estándar de la distribución normal que se considera en la hipótesis nula, usando los datos de la tabla 12.10. La media muestral $\bar{x}$ y la desviación estándar muestral s se usan como estimadores puntuales de la media y de la desviación estándar de la distribución normal. Los cálculos son los siguientes.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{3421}{50} = 68.42$$

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{5310.0369}{49}} = 10.41$$

Con estos datos se establecen las hipótesis siguientes acerca de la distribución de las puntuaciones de examen.

H_0: La población de las puntuaciones de examen tiene una distribución normal, con una media de 68.42 y desviación estándar de 10.41.

H_a: La población de las puntuaciones de examen no tiene una distribución normal, con media de 68.42 y desviación estándar de 10.41.

En la figura 12.2 se muestra esta distribución normal hipotética.

FIGURA 12.2 DISTRIBUCIÓN NORMAL HIPOTÉTICA DE PUNTUACIONES EN LA PRUEBA DE APTITUDES DE CHEMLINE

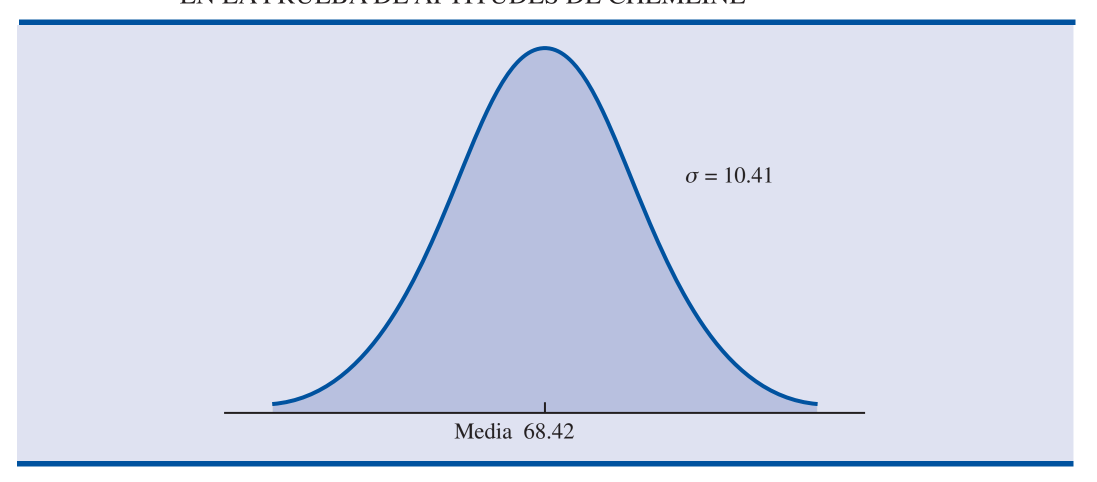
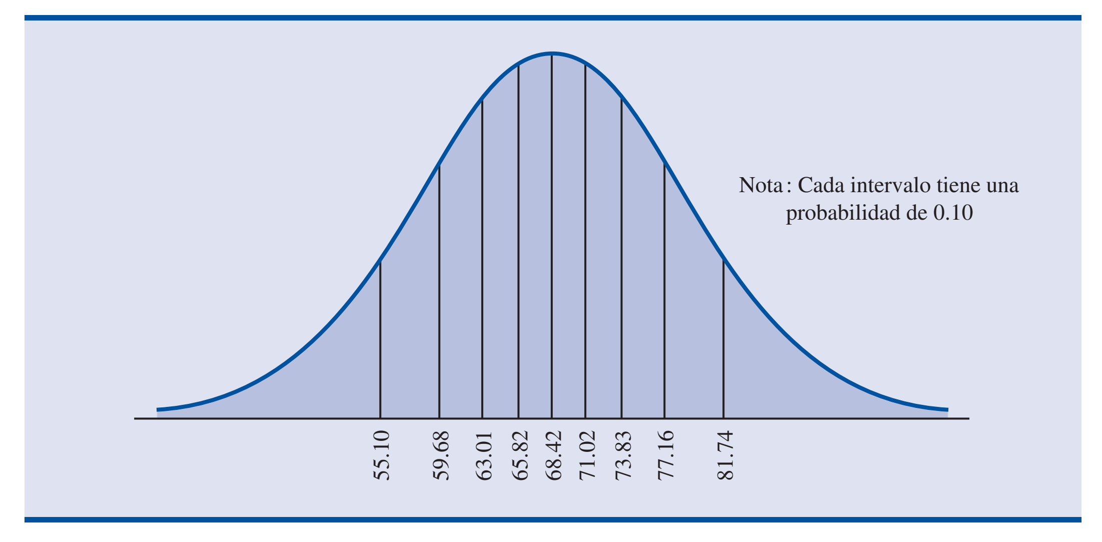




FIGURA 12.3 DISTRIBUCIÓN NORMAL EN EL EJEMPLO DE CHEMLINE CON 10 INTERVALOS DE PROBABILIDAD IGUAL



Ahora se verá cómo definir las categorías para una prueba de bondad de ajuste para una distribución normal. En el caso de la distribución de probabilidad discreta en la prueba para la distribución de Poisson, fue fácil definir las categorías en términos del número de llegadas de los clientes, 0, 1, 2, etc. Sin embargo, en el caso de la distribución de probabilidad normal que es continua, es necesario emplear un procedimiento diferente para definir las categorías. Se necesita definir las categorías en términos de *intervalos* de puntuaciones de examen.

Como se trata de una distribución de probabilidad continua, se establecen intervalos de manera que en cada uno la frecuencia esperada sea cinco o más.

Recuerde la regla de que en cada intervalo o categoría, la frecuencia esperada debe ser por lo menos cinco. Las categorías para las puntuaciones de examen se definen de manera que la frecuencia esperada en cada categoría sea por lo menos cinco. Como el tamaño de la muestra es 50, una manera de establecer las categorías es dividir la distribución normal en 10 intervalos con una misma probabilidad (véase la figura 12.3). Como el tamaño de la muestra es 50, se espera tener cinco resultados en cada intervalo o categoría, con lo que se satisface la regla de las cinco frecuencias esperadas.

El procedimiento para calcular los límites de las categorías es el siguiente. Como se trata de una distribución de probabilidad normal, para determinar estos límites se emplean las tablas de la distribución de probabilidad normal estándar. Primero se determina la puntuación de examen que separa el 10% inferior de las puntuaciones. En la tabla 1 del apéndice B se encuentra que el valor z correspondiente a esta puntuación de examen es -1.28. Por tanto, la puntuación de examen $x = 68.42 - 1.28(10.41) = 55.10$ es el valor que separa el 10% inferior de las puntuaciones de examen. Para el 20% inferior se tiene $z = -0.84$ y, por tanto, $x = 68.42 - 0.84(10.41) = 59.68$. Continuando de esta manera se obtienen los valores siguientes para las puntuaciones de examen.

Porcentaje	z	Puntuación de examen
10%	-1.28	$68.42 - 1.28(10.41) = 55.10$
20%	-0.84	$68.42 - 0.84(10.41) = 59.68$
30%	-0.52	$68.42 - 0.52(10.41) = 63.01$
40%	-0.25	$68.42 - 0.25(10.41) = 65.82$
50%	0.00	$68.42 + 0(10.41) = 68.42$
60%	$+0.25$	$68.42 + 0.25(10.41) = 71.02$
70%	$+0.52$	$68.42 + 0.52(10.41) = 73.83$
80%	$+0.84$	$68.42 + 0.84(10.41) = 77.16$
90%	$+1.28$	$68.42 + 1.28(10.41) = 81.74$

En la gráfica 12.3 se observan estos puntos de separación o límites de los intervalos.

TABLA 12.11 FRECUENCIAS ESPERADAS Y OBSERVADAS DE LAS PUNTUACIONES DE EXAMEN DE LOS SOLICITANTES DE EMPLEO EN CHEMLINE

Intervalo de puntuaciones de examen	Frecuencia observada (f_i)	Frecuencia esperada (e_i)
Menores que 55.10	5	5
55.10 a 59.68	5	5
59.68 a 63.01	9	5
63.01 a 65.82	6	5
65.82 a 68.42	2	5
68.42 a 71.02	5	5
71.02 a 73.83	2	5
73.83 a 77.16	5	5
77.16 a 81.74	5	5
81.74 o más	6	5
Total	50	50

Una vez definidas las categorías o intervalos de las puntuaciones de examen y siendo que la frecuencia esperada en cada categoría es cinco, se usan los datos muestrales de la tabla 12.10 y se determinan las frecuencias observadas en estas categorías. Con esto se obtienen los resultados que aparecen en la tabla 12.11.

Una vez que se tienen los resultados de la tabla 12.11, la prueba de bondad de ajuste procede exactamente como antes. Es decir, se comparan los resultados observados y esperados calculando el valor de χ^2. En la tabla 12.12 se muestran los cálculos necesarios para obtener el estadístico de prueba chi-cuadrada. Como se ve, el valor del estadístico de prueba es $\chi^2 = 7.2$.

Para determinar si este valor de 7.2 obtenido para χ^2 es suficientemente grande para rechazar H_0 se necesita consultar las tablas de la distribución chi-cuadrada. Al aplicar la regla para el cálculo del número de grados de libertad en la prueba de bondad de ajuste, se tiene, $k - p - 1 = 10 - 2 - 1 = 7$ grados de libertad, ya que se tienen 10 categorías y $p = 2$ parámetros (media y desviación estándar) estimados mediante los datos muestrales.

TABLA 12.12 CÁLCULO DEL ESTADÍSTICO DE PRUEBA CHI-CUADRADA EN EL EJEMPLO DE LAS PUNTUACIONES DE EXAMEN DE LOS SOLICITANTES DE EMPLEO EN CHEMLINE

Intervalos de puntuaciones de examen	Frecuencia observada (f_i)	Frecuencia esperada (e_i)	Diferencia ($f_i - e_i$)	Cuadrado de la diferencia $(f_i - e_i)^2$	Cuadrado de la diferencia dividido entre la frecuencia esperada $(f_i - e_i)^2/e_i$
Menores que 55.10	5	5	0	0	0.0
55.10 a 59.68	5	5	0	0	0.0
59.68 a 63.01	9	5	4	16	3.2
63.01 a 65.82	6	5	1	1	0.2
65.82 a 68.42	2	5	−3	9	1.8
68.42 a 71.02	5	5	0	0	0.0
71.02 a 73.83	2	5	−3	9	1.8
73.83 a 77.16	5	5	0	0	0.0
77.16 a 81.74	5	5	0	0	0.0
81.74 y más	6	5	1	1	0.2
Total	50	50			$\chi^2 = 7.2$

Como se estiman dos parámetros de la distribución normal, se pierden dos grados de libertad para la prueba χ^2.

Suponga que se prueba la hipótesis nula de que la distribución de las puntuaciones de examen es una distribución normal usando 0.10 como nivel de significancia. Para probar esta hipótesis se necesita determinar el valor-*p* para el estadístico de prueba $\chi^2 = 7.2$ determinando el área en la cola superior correspondiente en la distribución chi-cuadrada con 7 grados de libertad. Consultando la tabla 3 del apéndice B, se encuentra que el área en la cola superior correspondiente a $\chi^2 = 7.2$ es mayor que 0.10. Por consiguiente, se sabe que el valor-*p* es mayor que 0.10. Con los procedimientos de Minitab y Excel presentados en el apéndice F al final del libro, se encuentra que $\chi^2 = 7.2$ da un valor-$p = 0.4084$. Como el valor-$p > \alpha = 0.10$ no se puede rechazar la hipótesis nula de que las puntuaciones de examen de los solicitantes de empleo en Chemline sea una distribución normal. La distribución normal se puede usar como ayuda en la interpretación de las puntuaciones de examen. A continuación se presenta un resumen de la prueba de bondad de ajuste para una distribución normal.

PRUEBA DE BONDAD DE AJUSTE PARA UNA DISTRIBUCIÓN NORMAL: RESUMEN

1. Establecer las hipótesis nula y alternativa.

H_0: La población tiene una distribución normal
H_a: La población no tiene una distribución normal

2. Tomar una muestra aleatoria y
  a. Calcular la media muestral y la desviación estándar muestral.
  b. Definir intervalos de valores de manera que la frecuencia esperada en cada intervalo sea por lo menos cinco. Usar intervalos de igual probabilidad es un buen enfoque.
  c. En cada uno de los intervalos definidos anotar la frecuencia observada f_i en los datos.

3. Calcular el número esperado de ocurrencias e_i en cada uno de los intervalos de valores definidos en el paso 2 b. Multiplicar el tamaño de la muestra por la probabilidad de que una variable aleatoria normal pertenezca al intervalo.

4. Calcular el valor del estadístico de prueba

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

5. Regla de rechazo:

Método del valor-*p*: Rechazar H_0 si valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $\chi^2 \geq \chi^2_\alpha$

donde α es el nivel de significancia y los grados de libertad son $k - 3$.

Ejercicios

Métodos



20. A continuación se presenta el número de ocurrencias por lapso de tiempo y su frecuencia observada. Use $\alpha = 0.05$ y la prueba de bondad de ajuste para ver si estos datos se ajustan a una distribución de Poisson.

Número de ocurrencias	Frecuencia observada
0	39
1	30
2	30
3	18
4	3



21. Los datos siguientes provienen de una distribución normal. Use la prueba de bondad de ajuste con $\alpha = 0.05$ para probar tal suposición.

17	23	22	24	19	23	18	22	20	13	11	21	18	20	21
21	18	15	24	23	23	43	29	27	26	30	28	33	23	29

Aplicaciones

22. Al parecer el número de accidentes automovilísticos por día en una determinada ciudad tiene una distribución de Poisson. A continuación se presentan los datos de una muestra de 80 días del año anterior. ¿Estos datos apoyan la creencia de que el número de accidentes por día tiene una distribución de Poisson? Use $\alpha = 0.05$.

Número de accidentes	Frecuencia observada (días)
0	34
1	25
2	11
3	7
4	3

23. El número de llamadas telefónicas que llegan por minuto al conmutador de una empresa tiene una distribución de Poisson. Use $\alpha = 0.10$ y los datos siguientes para probar esta suposición.

Número de llamadas telefónicas que llegan por minuto	Frecuencia observada
0	15
1	31
2	20
3	15
4	13
5	4
6	2
Total	100

24. La demanda semanal de un producto tiene una distribución normal. Aplique una prueba de bondad de ajuste y los datos siguientes para probar esta suposición. Use $a = 0.10$. La media muestral es 24.5 y la desviación estándar es 3.

18	20	22	27	22
25	22	27	25	24
26	23	20	24	26
27	25	19	21	25
26	25	31	29	25
25	28	26	28	24

25. Use $\alpha = 0.01$ y realice una prueba de bondad de ajuste para ver si los datos siguientes parecen haber sido tomados de una distribución normal.

55	86	94	58	55	95	55	52	69	95	90	65	87	50	56
55	57	98	58	79	92	62	59	88	65					

Una vez terminada la prueba de bondad de ajuste, elabore un histograma con estos datos. ¿El histograma respalda la conclusión a la que se llegó con la prueba de bondad de ajuste? (*Nota:* $\bar{x} = 71$ y $s = 17$.)

Resumen

En este capítulo se presentó la prueba de bondad de ajuste y la prueba de independencia, las cuales se basan en el uso de la distribución chi-cuadrada. El objeto de la prueba de bondad de ajuste es determinar si una distribución de probabilidad hipotética sirve como modelo para una determinada población de interés. Al hacer los cálculos en una prueba de bondad de ajuste se comparan las frecuencias observadas en una muestra con las frecuencias esperadas suponiendo que la distribución de probabilidad hipotética sea verdadera. Para determinar si las diferencias entre frecuencias observadas y esperadas son suficientemente grandes para rechazar la distribución de probabilidad hipotética se usa la distribución chi-cuadrada. También se ilustró la prueba de bondad de ajuste para las distribuciones multinomial, de Poisson y normal.

Una prueba de independencia entre dos variables es una extensión de la metodología empleada en la prueba de bondad de ajuste para una población multinomial. Para determinar las frecuencias observadas y esperadas se emplea una tabla de contingencia. Después se calcula el valor de chi-cuadrada. Valores grandes de chi-cuadrada, debidos a diferencias grandes entre frecuencias observadas y esperadas, llevan al rechazo de la hipótesis nula de independencia.

Glosario

Población multinomial Población en la que cada elemento corresponde a una y sólo a una de varias categorías. Una distribución multinomial es una extensión de la distribución binomial a tres o más resultados.

Prueba de bondad de ajuste Prueba estadística que se realiza para determinar si se rechaza una distribución de probabilidad hipotética como distribución de una población.

Tabla de contingencia Tabla que se emplea para presentar las frecuencias observadas y esperadas en una prueba de independencia.

Fórmulas clave

Estadístico de prueba para la bondad de ajuste

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \tag{12.1}$$

Frecuencias esperadas para tablas de contingencia bajo la suposición de independencia

$$e_{ij} = \frac{(\text{Total del renglón } i)\ (\text{Total de la columna } j)}{\text{Tamaño de la muestra}} \tag{12.2}$$

Estadístico de prueba para independencia

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \tag{12.3}$$

Ejercicios complementarios

26. Para establecer cuotas de venta, el gerente de marketing supone que en los cuatro territorios de ventas el potencial de ventas es el mismo. A continuación se presenta una muestra de 200 ventas. ¿Debe rechazarse la suposición del gerente? Use $\alpha = 0.05$.

Territorios de venta			
I	**II**	**III**	**IV**
60	45	59	36

27. Siete por ciento de quienes invierten en fondos mutualistas consideran que las acciones corporativas son "muy seguras", 58% las considera "relativamente seguras", 24% las considera "no muy seguras", 4% las considera "no seguras" y 7% "no están seguros". *Business*/*Week*/Harris preguntó a 529 inversionistas de fondos mutualistas cómo calificarían ellos los bonos corporativos respecto de su seguridad. Las respuestas fueron las siguientes

Seguridad	Frecuencia
Muy seguros	48
Relativamente seguros	323
No muy seguros	79
Nada seguros	16
No están seguros	63
Total	529

¿La actitud de los inversionistas en fondos mutualistas difiere respecto a los bonos corporativos de su actitud frente a las acciones corporativas? Apoye su conclusión dando una prueba estadística. Use $\alpha = 0.01$

28. Desde el año 2000, Toyota Camry, Honda Accord y Ford Taurus han sido los tres automóviles de pasajeros en Estados Unidos mejor vendidas. Los datos de ventas de 2003 indican que las participaciones en el mercado de estos tres automóviles son las siguientes: Toyota Camry 37%, Honda Accord 34% y Ford Taurus 29%. Suponga que en una muestra de 1 200 ventas de automóviles de pasajeros durante el primer trimestre de 2004 se encuentran los datos siguientes.

Automóviles de pasajeros	Unidades vendidas
Toyota Camry	480
Honda Accord	390
Ford Taurus	330

¿Estos datos sirven para concluir que las participaciones en el mercado de estos tres automóviles de pasajeros cambiaron en el primer trimestre de 2004? ¿Cuál es el valor-*p*? Use 0.05 como nivel de significancia. ¿Cuál es su conclusión?

29. Una autoridad regional de tránsito está preocupada por el número de pasajeros en una de las rutas de autobús. Al establecer la ruta se supuso que el número de pasajeros era la misma todos los días de la semana, de lunes a viernes. Con los datos siguientes y usando $a = 0.05$ determine si la suposición de la autoridad de transito es correcta.

Día	Números de pasajeros
Lunes	13
Martes	16
Miércoles	28
Jueves	17
Viernes	16

30. *Computerworld's* Annual Job Satisfaction Survey encontró que 28% de los administradores de sistemas de información (SI) estaban muy satisfechos con su trabajo, 46% estaban moderadamente satisfechos con su trabajo, 12% no estaban ni satisfechos ni insatisfechos, 10% estaban ligeramente insatisfechos y 4% estaban muy insatisfechos. Suponga que en una muestra de 500 programadores se encontraron los resultados siguientes.

Categoría	Número de entrevistados
Muy satisfechos	105
Moderadamente satisfechos	235
Ni satisfechos ni insatisfechos	55
Ligeramente insatisfechos	90
Muy insatisfechos	15

Use $\alpha = 0.05$ y realice una prueba para determinar si la satisfacción con el trabajo entre los programadores de computadoras es diferente de la satisfacción con el trabajo de los administradores de SI.

31. De una muestra de piezas se obtiene la tabla de contingencia siguiente sobre la calidad, de acuerdo con el turno de producción.

Turno	Números de piezas	Números de defectuosos
Primero	368	32
Segundo	285	15
Tercero	176	24

Use $\alpha = 0.05$ para probar la hipótesis de que la calidad es independiente del turno de producción. ¿Cuál es la conclusión?

32. *The Wall Street Journal* hizo un estudio sobre el tipo de empleo de sus suscriptores. Los siguientes datos muestrales corresponden a las ediciones del este y del oeste.

	Región	
Tipo de empleo	**Edición del este**	**Edición del oeste**
Tiempo completo	1105	574
Medio tiempo	31	15
Autoempleo/consultor	229	186
No empleado	485	344

Use $\alpha = 0.05$ para probar la hipótesis de que el tipo de empleo es independiente de la región. ¿Cuál es su conclusión?

33. Una institución de préstamo muestra los datos siguientes sobre los préstamos aprobados por cuatro de sus agentes. Use $\alpha = 0.05$ y realice una prueba para determinar si la aprobación de las decisiones de préstamo es independiente del agente que recibe la solicitud de préstamo.

	Decisión de aprobar el préstamo	
Agente de préstamo	**Aprobada**	**Rechazada**
Miller	24	16
McMahon	17	13
Games	35	15
Runk	11	9

34. Como parte de un estudio nacional se obtuvieron datos sobre el estado civil de hombres y mujeres de 20 a 29 años. Los resultados en una muestra de 350 hombre y 400 mujeres son los siguientes.

	Estado civil		
Género	**Soltero**	**Casado**	**Divorciado**
Hombre	234	106	10
Mujer	216	168	16

a. Use $\alpha = 0.01$ para probar la independencia entre el estado civil y el género. ¿Cuál es su conclusión?

b. Dé el porcentaje en cada una de las categorías de estado civil de hombres y mujeres.

35. Barna Research Group presenta datos obtenidos sobre la asistencia a la iglesia de acuerdo con las edades (*USA Today*, 20 de noviembre de 2003). Use los datos muestrales para determinar si la asistencia a la iglesia es independiente de la edad. Use 0.05 como nivel de significancia. ¿Cuál es su conclusión? ¿Qué conclusión se puede sacar acerca de la asistencia a la iglesia a medida que las personas envejecen?

	Asistencia a la iglesia		
Edad	**Sí**	**No**	**Total**
20 a 29	31	69	100
30 a 39	63	87	150
40 a 49	94	106	200
50 a 59	72	78	150

36. Los siguientes son datos sobre el número de llamadas solicitando una ambulancia de emergencia en una zona rural y en una zona urbana de Virginia.

		Día de la semana							
		Domingo	**Lunes**	**Martes**	**Miércoles**	**Jueves**	**Viernes**	**Sábado**	**Total**
Zona	**Urbana**	61	48	50	55	63	73	43	393
	Rural	7	9	16	13	9	14	10	78
	Total	68	57	66	68	72	87	53	471

Realice una prueba de independencia usando $\alpha = 0.05$. ¿Cuál es su conclusión?

37. Las siguientes son las calificaciones en los exámenes finales en un curso universitario.

55	85	72	99	48	71	88	70	59	98	80	74	93	85	74
82	90	71	83	60	95	77	84	73	63	72	95	79	51	85
76	81	78	65	75	87	86	70	80	64					

Use $\alpha = 0.05$ y realice una prueba para determinar si se debe rechazar que una distribución normal sea representativa de la distribución poblacional de estas calificaciones.

38. Los datos siguientes dan el índice de ocupación de las oficinas en cuatro zonas metropolitanas de California. ¿Los datos indican que la cantidad de oficinas libres es independiente de la zona metropolitana? Use 0.05 como nivel de significancia. ¿Cuál es su conclusión?

Situación	Los Angeles	San Diego	San Francisco	San Jose
Ocupadas	160	116	192	174
Libres	40	34	33	26

39. Un vendedor hace cuatro llamadas por día. En una muestra de 100 días los volúmenes de venta son los siguientes.

Número de ventas	Frecuencia observada (días)
0	30
1	32
2	25
3	10
4	3
Total	100

Por experiencia se sabe que 30% de las llamadas llevan a una venta. Si las llamadas de ventas son independientes, el número de ventas por día deberá seguir una distribución binomial. La función de probabilidad binomial, presentada en el capítulo 5 es

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

En este ejercicio suponga que la población tiene una distribución binomial con $n = 4$, $p = 0.30$ y $x = 0, 1, 2, 3$ y 4.

a. Mediante la distribución de probabilidad binomial, calcule las frecuencias esperadas para $x = 0, 1, 2, 3$ y 4. Si es necesario combine categorías para satisfacer el requerimiento de que la frecuencia esperada en cada categoría debe ser cinco o más.
b. Use la prueba de bondad de ajuste para determinar si se debe rechazar la suposición de una distribución binomial. Use $\alpha = 0.05$. Como no hubo necesidad de estimar ninguno de los parámetros de la distribución binomial a partir de los datos muestrales, los grados de libertad son $k - 1$, donde k es el número de categorías.

Caso problema Una agenda bipartidista para el cambio

En un estudio realizado por Zogby International para *Democrat and Chronicle,* se entrevistaron más de 700 neoyorquinos para determinar si el gobierno del estado de Nueva York funcionaba. Entre los asuntos sobre los que se interrogaba a los entrevistados estaban reducciones de salario

a los legisladores, restricciones a los grupos de presión, límites de mandato para los legisladores y si los ciudadanos podrían incluir sus temas en las consultas ciudadanas (*Democrat and Chronicle*, 7 de diciembre de 1997). Los resultados mostraron un amplio apoyo a varias reformas, en niveles políticos y demográficos.

Suponga que en un estudio subsiguiente se entrevistan 100 individuos que viven en la región oeste de Nueva York. De cada entrevistado se registra su afiliación partidaria (demócrata, independiente o republicano), así como sus respuestas a estas tres preguntas.

1. ¿Se les debe reducir el sueldo a los legisladores por cada día que se retrasa el presupuesto para el estado?
   Sí___ No___
2. ¿Debe haber más restricciones para los grupos de presión?
   Sí___ No___
3. ¿Debe haber límites para que el mandato de los legisladores sea de un número determinado de años?
   Sí___ No___



Las respuestas fueron codificadas usando 1 para Sí y 2 para No. Los datos obtenidos se encuentran en el archivo titulado NYReform del disco compacto.

Informe administrativo

1. Use estadísticos descriptivos para resumir los datos de este estudio. ¿Cuáles son, respecto de cada pregunta, las conclusiones preliminares acerca de la independencia entre repuesta (Sí, No) y afiliación política?
2. Respecto de la pregunta 1, pruebe la independencia entre la respuesta (Sí, No) y afiliación política. Use $\alpha = 0.05$.
3. Respecto de la pregunta 2, pruebe la independencia entre la respuesta (Sí, No) y afiliación política. Use $\alpha = 0.05$.
4. Respecto de la pregunta 3, pruebe la independencia entre la respuesta (Sí, No) y afiliación política. Use $\alpha = 0.05$.
5. ¿Parece haber un amplio apoyo para los cambios en todos los estratos políticos? Explique.

Apéndice 12.1 Pruebas de bondad de ajuste e independencia mediante Minitab

Prueba de bondad de ajuste

Este procedimiento de Minitab se usa para pruebas de bondad de ajuste para la distribución multinomial de la sección 12.1 y las distribuciones de Poisson y normal de la sección 12.3. El usuario tendrá que obtener las frecuencias observadas, calcular las frecuencias esperadas e ingresar tanto frecuencias observadas como esperadas en la hoja de cálculo de Minitab. La columna C1 se rotula como Observada y contiene las frecuencias observadas. La columna C2 se rotula como Esperadas y contiene las frecuencias esperadas. Use el ejemplo de Scott Marketing Research de la sección 12.1, abra una hoja de cálculo de Minitab e ingrese las frecuencias observadas 48, 98 y 54 en la columna C1 y las frecuencias esperadas 60, 100 y 40 en la columna C2. Los pasos para la prueba de bondad de ajuste usando Minitab son los siguientes.

Paso 1. Seleccionar el menú **Calc**
Paso 2. Elegir **Calculator**

Paso 3. Cuando aparezca el cuadro de diálogo Calculator:
Ingresar ChiSquare en el cuadro **Store result in variable**
Ingresar Sum((C1-C2)**2/C2) en el cuadro **Expression**
Clic en **OK**
Paso 4. Seleccionar el menú **Calc**
Paso 5. Elegir **Probability Distributions**
Paso 6. Elegir **Chi-Square**
Paso 7. Cuando aparezca el cuadro de diálogo Chi-Square Distribution:
Seleccionar **Cumulative probability**
Ingresar 2 en el cuadro **Degree of freedom**
Seleccionar **Input column** e ingresar ChiSquare en el cuadro
Clic en **OK**

En los resultados que da Minitab presenta la probabilidad acumulada 0.9745, que es el área bajo la curva a la izquierda de $\chi^2 = 7.34$. El área restante en la cola superior es el valor-p. Por tanto, valor-$p = 1 - 0.9745 = 0.0255$.

Prueba de independencia

Con el ejemplo de la Albert's Brewery de la sección 12.2, se empieza con una nueva hoja de cálculo de Minitab y se ingresan los datos de las frecuencias observadas en las columnas 1, 2 y 3, respectivamente. Es decir, las frecuencias observadas que corresponden a las preferencias por la cerveza ligera (20 y 30) se ingresan en la columna C1, las frecuencias observadas que corresponden a las preferencias por la cerveza clara (40 y 30) se ingresan en la columna C2 y las frecuencias observadas que corresponden a las preferencias por la cerveza oscura (20 y 10) se ingresan en la columna C3. Los pasos para la prueba de independencia usando Minitab son los siguientes.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Seleccionar **Tables**
Paso 3. Elegir **Chi-Square Test (Table in Worksheet)**
Paso 4. Cuando aparezca el cuadro de diálogo Chi-Square Test
Ingresar C1-C3 en el cuadro **Columns containing the table**
Clic en **OK**

Apéndice 12.2 Pruebas de bondad de ajuste e independencia mediante Excel

Prueba de bondad de ajuste



FitTest

Este procedimiento de Excel se usa para pruebas de bondad de ajuste para la distribución multinomial de la sección 12.1 y las distribuciones de Poisson y normal de la sección 12.3. El usuario tendrá que obtener las frecuencias observadas, calcular las frecuencias esperadas e ingresar tanto frecuencias observadas como esperadas en la hoja de cálculo de Excel.

Las frecuencias observadas y las esperadas en el ejemplo de Scott Market Research de la sección 12.1 se ingresan en las columnas A y B, como se muestra en la figura 12.4. El estadístico de prueba $\chi^2 = 7.34$ se calcula en la columna D. Como hay $k = 3$ categorías, el usuario ingresa los grados de libertad $k - 1 = 3 - 1 = 2$ en la celda D11. La función CHIDISTR.CHI. proporciona el valor-p en la celda D13. En la hoja de cálculo que aparece en segundo plano se presentan las fórmulas correspondientes a cada celda.

FIGURA 12.4 HOJA DE CÁLCULO DE EXCEL PARA LA PRUEBA DE BONDAD DE AJUSTE EN EL EJEMPLO DE SCOTT MARKET RESEARCH

	A	B	C	D	E
1	**Goodness of Fit Test**				
2					
3	Observed	Expected			
4	Frequency	Frequency		Calculations	
5	48	60		=(A5-B5)^2/B5	
6	98	100		=(A6-B6)^2/B6	
7	54	40		=(A7-B7)^2/B7	
8					
9			Test Statistic	=SUM(D5:D7)	
10					
11			Degrees of Freedom	2	
12					
13			p-Value	=CHIDIST(D9,D11)	
14					

	A	B	C	D	E
1	**Goodness of Fit Test**				
2					
3	Observed	Expected			
4	Frequency	Frequency		Calculations	
5	48	60		2.40	
6	98	100		0.04	
7	54	40		4.90	
8					
9			Test Statistic	7.34	
10					
11			Degrees of Freedom	2	
12					
13			p-Value	0.0255	
14					

Prueba de independencia



Independence

En el procedimiento de Excel para pruebas de independencia se requiere que el usuario obtenga las frecuencias observadas y las ingrese en una hoja de cálculo. En el ejemplo de la Alber's Brewery presentado en la sección 12.2 se dan las frecuencias observadas, las cuales se ingresan en las celdas B7 a D8, como se muestra en la hoja de cálculo de la figura 12.5. Las fórmulas que aparecen en las celdas de la hoja de cálculo en segundo plano muestran el procedimiento empleado para calcular las frecuencias esperadas. En la celda E22, se ingresan los grados de libertad, que como se tienen dos renglones y tres columnas, serán $(2 - 1)(3 - 1) = 2$. La función PRUEBA.CHI proporciona en la celda E24 el valor-p.

FIGURA 12.5 HOJA DE CÁLCULO DE EXCEL PARA LA PRUEBA DE INDEPENDENCIA DE LA ALBER'S BREWERY

	A	B	C	D	E	F
1	**Test of Independence**					
2						
3	**Observed Frequencies**					
4						
5			**Beer Peference**			
6	**Gender**	Light	Regular	Dark	Total	
7	Male	20	40	20	=SUM(B7:D7)	
8	Female	30	30	10	=SUM(B8:D8)	
9	Total	=SUM(B7:B8)	=SUM(C7:C8)	=SUM(D7:D8)	=SUM(E7:E8)	
10						
11						
12	**Expected Frequencies**					
13						
14			**Beer Peference**			
15	**Gender**	Light	Regular	Dark	Total	
16	Male	=E7*B$9/$E$9	=E7*C$9/$E$9	=E7*D$9/$E$9	=SUM(B16:D16)	
17	Female	=E8*B$9/$E$9	=E8*C$9/$E$9	=E8*D$9/$E$9	=SUM(B17:D17)	
18	Total	=SUM(B16:B17)	=SUM(C16:C17)	=SUM(D16:D17)	=SUM(E16:E17)	
19						
20				**Test Statistic**	=CHIINV(E24,E22)	
21						
22				**Degrees of Freedom**	2	
23						
24				**p-value**	=CHITEST(B7:D8,B16:D17)	
25						

	A	B	C	D	E	F
1	**Test of Independence**					
2						
3	**Observed Frequencies**					
4						
5			**Beer Peference**			
6	**Gender**	Light	Regular	Dark	Total	
7	Male	20	40	20	80	
8	Female	30	30	10	70	
9	Total	50	70	30	150	
10						
11						
12	**Expected Frequencies**					
13						
14			**Beer Peference**			
15	**Gender**	Light	Regular	Dark	Total	
16	Male	26.67	37.33	16	80	
17	Female	23.33	32.67	14	70	
18	Total	50	70	30	150	
19						
20				**Test Statistic**	6.12	
21						
22				**Degrees of Freedom**	2	
23						
24				**p-value**	0.0468	
25						



CAPÍTULO 13

Diseño de experimentos y análisis de varianza

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

BURKE MARKETING SERVICES, INC.*
CINCINNATI, OHIO

Burke Marketing Services, Inc., es una de las empresas de investigación de mercado con más experiencia. Cada día Burke presenta más propuestas, sobre más proyectos, que cualquier otra empresa de investigación de mercado en el mundo. Apoyada con la última tecnología, Burke ofrece una amplia variedad de posibilidades de investigación, con lo que da solución a casi cualquier problema de marketing.

En un estudio reciente una empresa solicitó los servicios de Burke para evaluar una nueva versión de un cereal para niños. Por razones de confidencialidad aquí se nombrará a ésta como empresa Anon. La empresa Anon consideraba que los cuatro factores principales que intervenían en el sabor del cereal eran

1. La proporción entre trigo y maíz en el cereal.
2. El tipo de edulcorante: azúcar, miel o edulcorante artificial.
3. La presencia o ausencia de trocitos con sabor a fruta.
4. El tiempo de cocción, largo o corto.

Burke diseñó un experimento para determinar el efecto de estos cuatro factores en el sabor del cereal. Por ejemplo, hizo una prueba con una determinada proporción de trigo y maíz, con azúcar como edulcorante, trocitos de sabor a fruta y tiempo corto de cocción; hizo otra prueba en la cual varió únicamente la proporción de trigo y maíz y dejó igual todos los demás factores, y así sucesivamente. Después un grupo de niños probó los cereales y dio su opinión acerca del sabor de cada uno.

El método estadístico empleado para estudiar los datos obtenidos de las pruebas de degustación fue el análisis de varianza. De los resultados de los análisis se concluyó lo siguiente:

- La relación entre trigo y maíz y el tipo de edulcorante influyeron de manera importante en la evaluación del sabor.
- Los trocitos con sabor a fruta, en realidad tuvieron un efecto negativo sobre el sabor del cereal.
- El tiempo de cocción no tuvo ninguna influencia sobre el sabor.

Con esta información Anon pudo identificar los factores que intervenían en la obtención del sabor del cereal.

El diseño experimental empleado por Burke y el posterior análisis de varianza sirvieron para hacer una recomendación en el diseño del producto. En este capítulo se verá cómo se realizan estos procedimientos.



Burke emplea pruebas de degustación para obtener información de lo que los clientes esperan de un producto. ©JLP/Sylvia Torres/CORBIS.

*Los autores agradecen a doctor Ronald Tatham de Burke Marketing Services por proporcionar este artículo para *La estadística en la práctica.*

En el capítulo 1 se dijo que los estudios estadísticos se clasifican como experimentales u observacionales. En un estudio estadístico experimental se realiza un experimento para obtener los datos. Un experimento empieza por la identificación de la variable de interés. Después se identifican y controlan una o más variables que se consideran relacionadas con la variable de interés, para después recoger datos de la influencia de estas variables sobre la variable de interés.

En un estudio observacional, los datos suelen obtenerse mediante inspección de una muestra y no mediante un experimento controlado. En estos estudios, aunque también se emplean los principios para un buen estudio, no es posible tener el control riguroso que se tiene en un estudio experimental. Por ejemplo, en un estudio para entender la relación entre fumar y el cáncer de pulmón, el investigador no puede asignarle a un sujeto el hábito de fumar. El investigador sólo puede observar los efectos de fumar en las personas que ya tienen este hábito y los efectos de no fumar en las personas que no lo tienen.

A sir Ronald Alymer Fisher (1890-1962) se le atribuye la invención de la rama de la estadística conocida como diseño de experimentos. Además de sus aportaciones a la estadística, fue un científico sobresaliente en el campo de la genética.

En este capítulo se presentan tres tipos de diseños de experimentos: un diseño completamente aleatorizado, un diseño de bloques aleatorizado y un experimento factorial. Para cada tipo de diseño se indica cómo usar el procedimiento estadístico conocido como análisis de varianza (ANOVA, por sus siglas en inglés) para analizar los datos de una variable. El ANOVA también se usa para analizar los datos obtenidos mediante un estudio observacional. Por ejemplo, se verá que el ANOVA también se usa en los diseños completamente aleatorizados para probar la igualdad de tres o más medias poblacionales de datos obtenidos mediante un estudio observacional. En los capítulos siguientes se verá que, además, el ANOVA tiene gran importancia en el análisis de los resultados de estudios de regresión, tanto de datos experimentales como observacionales.

En la primera sección de este capítulo se presentan los principios de un estudio experimental y cómo emplearlos en un diseño completamente aleatorizado. En la segunda sección se muestra cómo usar el ANOVA para analizar los datos de un diseño de experimentos completamente aleatorizado. En la última sección se estudian métodos de comparación múltiple y otros dos diseños de experimentos muy usados, el diseño de bloques aleatorizado y el experimento factorial.



Introducción al diseño de experimentos y al análisis de varianza

Las relaciones de causa y efecto son difíciles de establecer en estudios observacionales, pero fáciles de establecer en estudios experimentales.

Un ejemplo de un estudio estadístico experimental es el problema que se le presentó a la empresa Chemitech, Inc.; dicha empresa elaboró un sistema de filtración para los suministros de aguas municipales. Los componentes del sistema de filtración se comprarían a varios proveedores y Chemitech armaría el sistema de filtración en su fábrica en Columbia, Carolina del Sur. El grupo de ingenieros industriales era el encargado de determinar el mejor método para armar el sistema de filtración. Después de considerar varios métodos, quedaron sólo tres alternativas: el método A, el método B y el método C. La diferencia entre estos métodos era el orden en los pasos para armar el sistema. Los administradores de Chemitech, querían saber con qué método se podían producir más sistemas en una semana.

En el experimento de Chemitech, el método para armar el sistema es la variable independiente o **factor**. Como a este factor le corresponden tres métodos para armar el sistema, se dice que en este experimento hay tres **tratamientos**; cada tratamiento corresponde a uno de los tres métodos para armar el sistema. El problema de Chemitech es un ejemplo de un **experimento de un solo factor**; interviene sólo un factor cualitativo (el método para armar el sistema). En experimentos más complejos caben múltiples factores; los factores pueden ser cualitativos o cuantitativos.

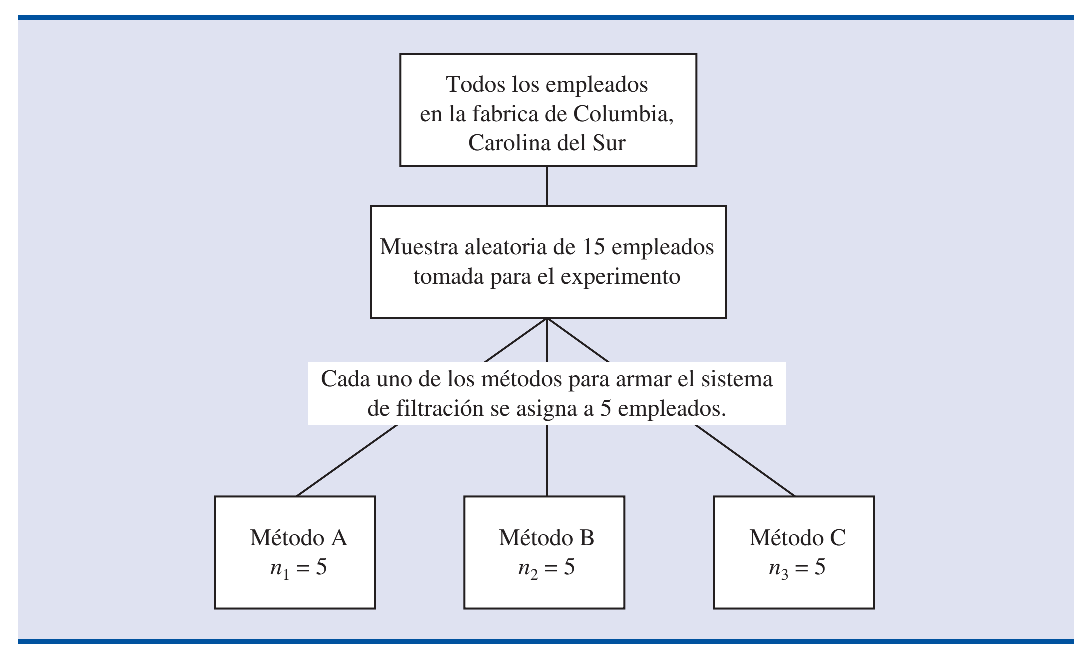
Aleatorización es el procedimiento por el cual se asignan al azar los tratamientos a las unidades experimentales. Antes del trabajo de sir R. A. Fisher, los tratamientos se asignaban de manera sistemática o subjetiva.

Los tres tratamientos o métodos para armar el sistema constituyen las tres poblaciones de interés del experimento de Chemitech. Una población está formada por todos los trabajadores que emplean el método A, otra población es la de todos los trabajadores que emplean el método B y otra población es la de todos los trabajadores que emplean el método C. Observe que en cada población la variable dependiente o **variable de respuesta** es el número de sistemas de filtración que se arman por semana, y el objetivo estadístico del experimento es determinar si el número medio producido por semana es el mismo en las tres poblaciones (con los tres métodos).

Suponga una muestra aleatoria de tres trabajadores de la empresa Chemitech. En el lenguaje del diseño de experimentos, estos tres trabajadores son las **unidades experimentales**. Al diseño de experimentos que se usará para el problema de Chemitech se le llama **diseño completamente aleatorizado**. En este tipo de diseño se requiere que cada uno de los tratamientos o métodos para armar el sistema se asigne de manera aleatoria a cada una de las unidades experimentales o trabajadores. Así, el método A le puede ser asignado aleatoriamente al segundo trabajador, el método B al primer trabajador y el método C al tercer trabajador. El concepto de *aleatorización*, como se ha ilustrado en este ejemplo, es importante en el diseño de experimentos.



FIGURA 13.1 DISEÑO COMPLETAMENTE ALEATORIZADO PARA EVALUAR EL MÉTODO EXPERIMENTAL DE ARMAR EL SISTEMA DE CHEMITECH



Observe que en este experimento sólo se obtendrá una medición (un dato) para cada método de armar el sistema de filtración. Para obtener más datos para cada método, se necesita repetir o replicar el proceso experimental básico. Considere que en lugar de tomar al azar sólo a tres trabajadores, se toman 15 trabajadores, y a cada cinco trabajadores se les asigna en forma aleatoria uno de los métodos para armar el sistema de filtración. Como cada uno de estos métodos es asignado a cinco trabajadores, se dice que se obtienen cinco réplicas. El proceso de *replicación* es otro principio importante en el diseño de experimentos. En la figura 13.1 se presenta el diseño completamente aleatorizado para el experimento de Chemitech.

Obtención de datos

Una vez satisfechos con el diseño del experimento, se procede a obtener y analizar los datos. En el caso de Chemitech, se les explicará a los trabajadores cómo emplear el método que les ha sido asignado y empezarán a armar los sistemas de filtración con ese método. En la tabla 13.1 se presenta el número de unidades armadas por cada empleado en una semana. En esta tabla se dan también la media muestral, la varianza muestral y la desviación estándar muestral obtenidas con cada método de ensamblado. Así, la media muestral del número de unidades producidas con el método A es 62; la media muestral con el método B es 66 y la media muestral usando el método C es 52. De acuerdo con estos datos, parece que con el método B se obtienen más unidades por semana que con los otros dos métodos.

Lo que importa es si las tres medias muestrales observadas difieren lo suficiente para poder concluir que las medias de las poblaciones correspondientes a estos tres métodos son diferentes. Para expresar esto en términos estadísticos se introduce la notación siguiente.

μ_1 = número promedio de unidades producidas por semana con el método A
μ_2 = número promedio de unidades producidas por semana con el método B
μ_3 = número promedio de unidades producidas por semana con el método C

TABLA 13.1 NÚMERO DE UNIDADES PRODUCIDA POR 15 TRABAJADORES



	Método		
	A	**B**	**C**
	58	58	48
	64	69	57
	55	71	59
	66	64	47
	67	68	49
Media muestral	62	66	52
Varianza muestral	27.5	26.5	31.0
Desviación estándar muestral	5.244	5.148	5.568

Aunque nunca se podrá saber cuáles son los verdaderos valores de μ_1, μ_2 y μ_3, se van a usar las medias muestrales para probar las hipótesis siguientes.

Si se rechaza H_0, no se puede concluir que todas las medias poblacionales sean diferentes. Rechazar H_0 significa que por lo menos dos de las medias poblacionales tienen un valor diferente.

$$H_0: \mu_1 = \mu_2 = \mu_3$$
$$H_a: \text{No todas las medias poblacionales son iguales}$$

Como se demostrará más adelante, el análisis de varianza (ANOVA) es el procedimiento estadístico que se emplea para determinar si las diferencias observadas entre las tres medias muestrales son lo suficientemente grandes para rechazar H_0.

Suposiciones para el análisis de varianza

Si los tamaños de las muestras son iguales, el análisis de varianza no es sensible a desviaciones de la suposición de que las poblaciones están distribuidas de manera normal.

Son tres las suposiciones para emplear el análisis de varianza.

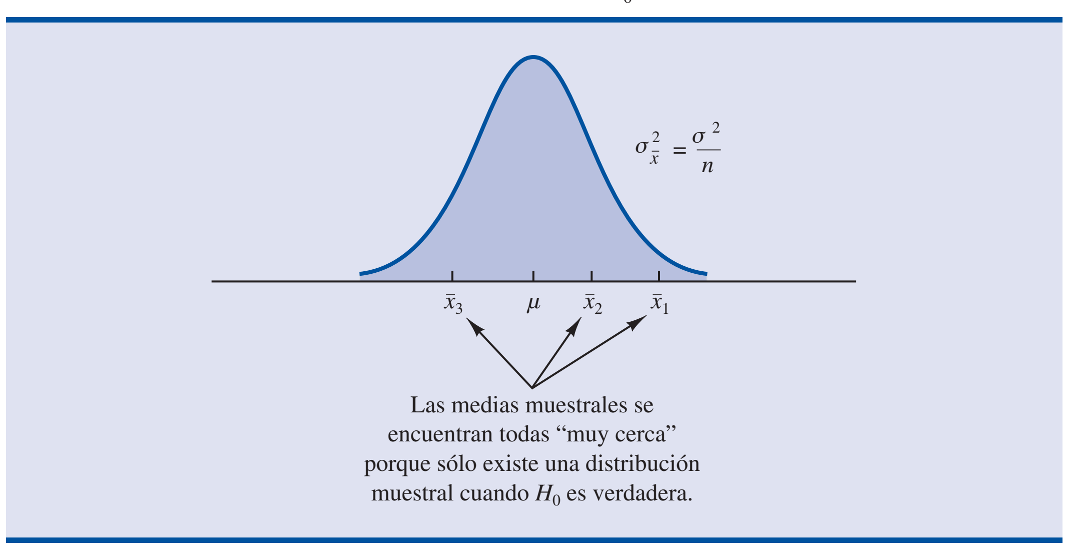
1. **En cada población, la variable de respuesta tiene una distribución normal.** Por tanto: en el experimento de Chemitech el número de unidades producida por semana (variable de respuesta) con cada uno de los métodos debe estar distribuida en forma normal.
2. **La varianza de la variable de respuesta, que se denota σ^2, es la misma en todas las poblaciones.** Por tanto: en el experimento de Chemitech, con los tres métodos, la varianza en el número de unidades producida por semana debe ser la misma.
3. **Las observaciones deben ser independientes.** Por tanto: en el experimento de Chemitech la cantidad de unidades producida por semana por un empleado debe ser independiente el número de unidades producidas por semana por cualquier otro empleado.

Análisis de varianza: una visión conceptual general

Si las medias de las tres poblaciones son iguales, se esperaría que las tres medias muestrales fueran muy parecidas. En efecto, entre más parecidas sean las medias muestrales, mayor será la evidencia para concluir que las medias poblacionales son iguales, o entre mayor sea la diferencia entre las medias muestrales, mayor será la evidencia para concluir que las medias poblacionales no son iguales. En otras palabras, si la variabilidad entre las medias muestrales es “pequeña”, esto favorece a H_0; si la variabilidad entre las medias muestrales es “grande”, esto favorece a H_a.

Si la hipótesis nula es verdadera, H_0: $\mu_1 = \mu_2 = \mu_3$, se usa la variabilidad entre las medias muestrales para estimar σ^2. Primero, observe que si se satisfacen las suposiciones para el análi-

FIGURA 13.2 DISTRIBUCIÓN MUESTRAL DE $\bar{x}$ SI H_0 ES VERDADERA



sis de varianza, cada una de las muestras provendrá de la misma distribución normal con media μ y varianza σ^2. Recuerde que en el capítulo 7 se vio que la distribución muestral de la media muestral $\bar{x}$ de una muestra aleatoria simple de tamaño n tomada de una población normal tendrá una distribución normal con media μ y desviación estándar σ^2/n. En la figura 13.2 se ilustra una distribución muestral así.

Por tanto, si la hipótesis nula es verdadera, se considera cada una de las tres medias muestrales, $\bar{x}_1 = 62$, $\bar{x}_2 = 66$ y $\bar{x}_3 = 52$ como valores obtenidos aleatoriamente de la distribución muestral que aparece en la figura 13.2. En este caso la media y la varianza de los tres valores $\bar{x}$ se usa para estimar la media y la varianza de la distribución muestral. Si los tamaños de las muestras son iguales, como en el caso de Chemitech, la mejor estimación de la media de la distribución muestral de $\bar{x}$ es la media o el promedio de las medias muestrales. Por tanto, en el experimento de Chemitech, una estimación de la media de la distribución muestral de $\bar{x}$ es $(62 + 66 + 52)/3 = 60$. A esta estimación se le conoce como *media muestral general*. Una estimación de la varianza de la distribución muestral de $\bar{x}$, $\sigma_{\bar{x}}^2$, se obtiene de la varianza de las tres medias muestrales.

$$s_{\bar{x}}^2 = \frac{(62 - 60)^2 + (66 - 60)^2 + (52 - 60)^2}{3 - 1} = \frac{104}{2} = 52$$

Como $\sigma_{\bar{x}}^2 = \sigma^2/n$, despejando σ^2 se obtiene

$$\sigma^2 = n\sigma_{\bar{x}}^2$$

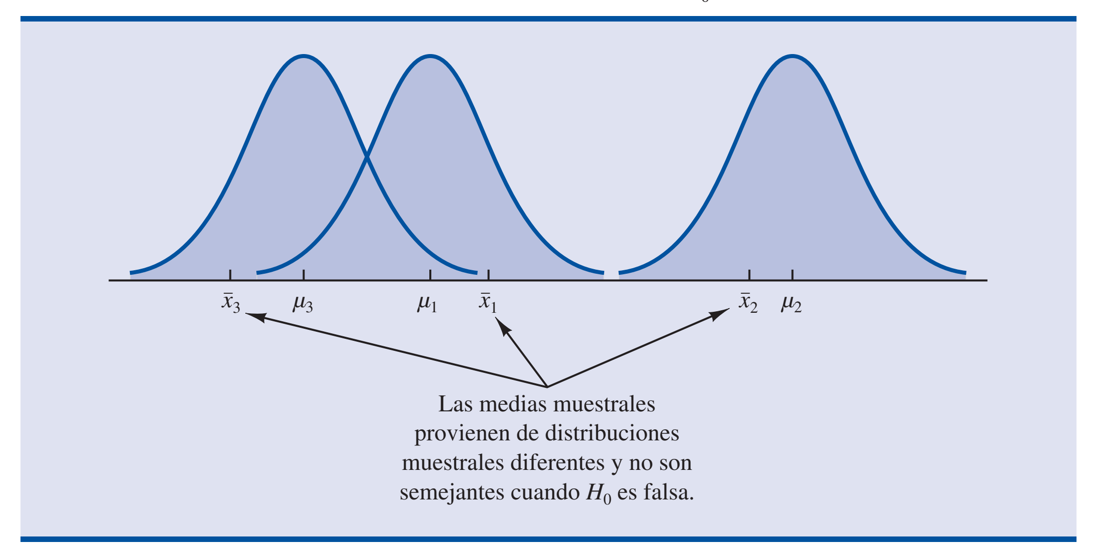
Por tanto,

$$\text{Estimación de } \sigma^2 = \text{n (Estimación de } \sigma_{\bar{x}}^2) = ns_{\bar{x}}^2 = 5(52) = 260$$

A $ns_{\bar{x}}^2 = 260$, se le conoce como estimación de σ^2 *entre tratamientos*.

La estimación de σ^2 entre tratamientos se basa en la suposición de que la hipótesis nula sea verdadera. En este caso cada una de las muestras proviene de la misma población y sólo hay una

FIGURA 13.3 DISTRIBUCIONES MUESTRALES DE $\bar{x}$ SI H_0 ES FALSA



distribución muestral de $\bar{x}$. Para ilustrar lo que ocurre cuando H_0 es falsa, suponga que las medias poblacionales son diferentes. Observe que como las tres muestras provienen de poblaciones normales con medias diferentes, darán tres distribuciones muestrales diferentes. En la figura 13.3 se muestra que en este caso las medias muestrales no están tan cerca unas de otras, como cuando la H_0 es verdadera. Entonces, $s_{\bar{x}}^2$ será mayor, haciendo que la estimación de σ^2 sea mayor. En general, cuando las medias poblacionales no son iguales, la estimación entre tratamientos sobreestimará la varianza poblacional σ^2.

La variación dentro de cada una de las muestras también tiene efecto sobre la conclusión a la que se arriba con el análisis de varianza. Cuando se toma una muestra aleatoria simple de cada población, cada una de las varianzas muestrales proporciona un estimador insesgado de σ^2. Por tanto, se combinan o juntan las estimaciones individuales de σ^2 en una estimación general. A la estimación de σ^2 obtenida de esta manera se le conoce como estimación *conjunta* o *dentro de los tratamientos* de σ^2. Como cada varianza muestral proporciona una estimación de σ^2 que se basa sólo en la variación dentro de cada muestra, a la estimación de σ^2 dentro de los tratamientos no le afecta que las medias poblacionales sean o no iguales.

Si los tamaños de las muestras son iguales, la estimación dentro de los tratamientos de σ^2 se obtiene del promedio de las varianzas muestrales. En el experimento de Chemitech se obtiene

$$\text{Estimación de } \sigma^2 \text{ dentro de los tratamientos} = \frac{27.5 + 26.5 + 31.0}{3} = \frac{85}{3} = 28.33$$

En el experimento de Chemitech, la estimación de σ^2 entre los tratamientos (260) es mucho mayor que la estimación de σ^2 dentro de los tratamientos (28.33). El cociente entre estas dos estimaciones es $260/28.33 = 9.18$. Pero debe recordarse que el método entre tratamientos sólo proporciona una buena estimación de σ^2 si la hipótesis nula es verdadera. Si la hipótesis nula es falsa, el método entre tratamientos sobreestima σ^2. El método dentro de los tratamientos proporciona una buena estimación de σ^2 en cualquiera de los casos. Por tanto si la hipótesis nula es verdadera, las dos estimaciones serán semejantes y su cociente será cercano a 1. Si la hipótesis nula es falsa, la estimación entre tratamientos será mayor que la estimación dentro de los tratamientos y su cociente será grande. En la sección siguiente se muestra cuán grande debe ser este cociente para que se rechace H_0.

En resumen, la idea detrás del ANOVA se basa en la obtención de dos estimaciones independientes de la varianza poblacional común σ^2. Una estimación de σ^2 se basa en la variabilidad entre las medias muestrales mismas y la otra estimación de σ^2 se basa en la variabilidad entre los datos dentro de cada muestra. Al comparar estas dos estimaciones de σ^2, se determina si las medias poblacionales son iguales.

NOTAS Y COMENTARIOS

1. En el diseño de experimentos, la aleatorización es el análogo al muestreo probabilístico en un estudio observacional.
2. En muchos experimentos médicos los sesgos potenciales se eliminan con el empleo de un diseño de experimento doble ciego. En este diseño, ni el médico ni el paciente saben qué tratamiento se está aplicando. Este tipo de diseño también es útil en muchos otros tipos de experimentos.
3. En esta sección se presentó una visión conceptual del uso del análisis de varianza para probar la igualdad de k medias poblacionales en un diseño experimental completamente aleatorizado. Se verá que este mismo procedimiento también se usa para probar la igualdad de k medias poblacionales en un estudio observacional o no experimental.
4. En las secciones 10.1 y 10.2 se presentaron métodos estadísticos para probar las hipótesis de igualdad de dos medias poblacionales El ANOVA también puede usarse para probar las hipótesis de que las medias de dos poblaciones son iguales. Sin embargo, en la práctica el análisis de varianza se usa cuando se tienen tres o más medias poblacionales



Análisis de varianza y el diseño completamente aleatorizado

En esta sección se muestra el uso del análisis de varianza para probar la igualdad de k medias poblacionales en un diseño completamente aleatorizado. La forma general de esta prueba de hipótesis es

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a: \text{No todas las medias poblacionales son iguales}$$

donde

$$\mu_j = \text{media de la } j\text{-ésima población}$$

Se supone que de cada una de las k poblaciones o tratamientos se toma una muestra aleatoria simple de tamaño n_j. Para los datos muestrales, sean

$$x_{ij} = \text{valor de la observación } i \text{ del tratamiento } j$$
$$n_j = \text{número de observaciones en el tratamiento } j$$
$$\bar{x}_j = \text{media muestral del tratamiento } j$$
$$s_j^2 = \text{varianza muestral del tratamiento } j$$
$$s_j = \text{desviación estándar muestral del tratamiento } j$$

Las fórmulas para la media muestral y la varianza muestral del tratamiento j son las siguientes:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \tag{13.1}$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1} \tag{13.2}$$

La media muestral general que se denota $\bar{\bar{x}}$, es la suma de todas las observaciones dividida entre la cantidad total de todas las observaciones. Es decir,

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T} \tag{13.3}$$

donde

$$n_T = n_1 + n_2 + \cdots + n_k \tag{13.4}$$

Si todas las muestras son de tamaño n, $n_T = kn$; en este caso, la ecuación 13.3 se reduce a

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{kn} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}/n}{k} = \frac{\sum_{j=1}^{k} \bar{x}_j}{k} \tag{13.5}$$

En otras palabras, si todas las muestras son del mismo tamaño, la media muestral general es el promedio de las k medias muestrales.

En el experimento de Chemitech, como todas las muestras constaban de $n = 5$ observaciones, la media muestral general se puede calcular empleando la fórmula 13.5. De acuerdo con los datos de la tabla 13.1 se tiene lo siguiente.

$$\bar{\bar{x}} = \frac{62 + 66 + 52}{3} = 60$$

Si la hipótesis nula es verdadera ($\mu_1 = \mu_2 = \mu_3 = \mu$), la media muestral general, 60, es la mejor estimación de la media poblacional μ.

Estimación de la varianza poblacional entre tratamientos

En la sección anterior se presentó el concepto de estimación de σ^2 entre tratamientos y se mostró cómo calcular esta estimación cuando todas las muestras eran del mismo tamaño. A esta estimación de σ^2 se le llama *cuadrado medio debido a los tratamientos* y se denota CMTR. La fórmula general para calcular el CMTR es

$$\text{CMTR} = \frac{\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2}{k - 1} \tag{13.6}$$

Al numerador de la ecuación (13.6) se le llama *suma de cuadrados debido a los tratamientos* y se denota (SCTR). El denominador, $k - 1$, representa los grados de libertad que corresponden a la SCTR. Por tanto, el cuadrado medio debido a los tratamientos se calcula mediante las fórmulas siguientes.

CUADRADO MEDIO DEBIDO A LOS TRATAMIENTOS

$$\text{CMTR} = \frac{\text{SCTR}}{k - 1} \tag{13.7}$$

donde

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 \tag{13.8}$$

Si H_0 es verdadera, el CMTR proporciona una estimación insesgada de σ^2. Pero, si las medias de las k poblaciones no son iguales, el CMTR no es un estimador insesgado de σ^2; en este caso el CMTR sobreestima σ^2.

En el caso de Chemitech, de acuerdo con los datos de la tabla 13.1, se tiene:

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 5(62 - 60)^2 + 5(66 - 60)^2 + 5(52 - 60)^2 = 520$$

$$\text{CMTR} = \frac{\text{SCTR}}{k - 1} = \frac{520}{2} = 260$$

Estimación de la varianza poblacional dentro de los tratamientos

En párrafos anteriores ya se presentó el concepto de estimación de σ^2 dentro de los tratamientos y se mostró cómo calcular esta estimación cuando todas las muestras son del mismo tamaño. A esta estimación de σ^2 se le llama *cuadrado medio debido al error* y se denota CME. La fórmula general para calcular el CME es

$$\text{CME} = \frac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{n_T - k} \tag{13.9}$$

Al numerador de la ecuación (13.9) se le llama *suma de cuadrados debido al error* y se denota SCE. El denominador del CME son los grados de libertad correspondientes a SCE. Por tanto, la fórmula para el CME también se expresa como sigue.

CUADRADO MEDIO DEBIDO AL ERROR

$$\text{CME} = \frac{\text{SCE}}{n_T - k} \tag{13.10}$$

donde

$$\text{SCE} = \sum_{j=1}^{k}(n_j - 1)s_j^2 \tag{13.11}$$

Observe que el CME está basado en la variación dentro de cada tratamiento; el que la hipótesis nula sea o no verdadera no tiene ninguna influencia. Por tanto, el CME siempre proporciona una estimación insesgada de σ^2.

De acuerdo con los datos de la tabla 13.1 para el caso de Chemitech, se obtienen los resultados siguientes.

$$\text{SCE} = \sum_{j=1}^{k}(n_j - 1)s_j^2 = (5-1)27.5 + (5-1)26.5 + (5-1)31 = 340$$

$$\text{CME} = \frac{\text{SCE}}{n_T - k} = \frac{340}{15-3} = \frac{340}{12} = 28.33$$

Comparación de las estimaciones de las varianzas: la prueba *F*

En la sección 11.2 se hizo una introducción a la distribución F y al uso de las tablas de la distribución F.

Si la hipótesis nula es verdadera, el CMTR y el CME proporcionan dos estimaciones insesgadas e independientes de σ^2. De acuerdo con lo visto en el capítulo 11, cuando se tienen poblaciones normales, la distribución muestral del cociente de dos estimaciones independientes de σ^2 sigue una distribución F. Por tanto, si la hipótesis nula es verdadera y si se satisfacen las suposiciones del ANOVA, la distribución muestral de CMTR/CME será una distribución F con $k - 1$ grados de libertad en el numerador y $n_T - k$ grados de libertad en el denominador. En otras palabras, si la hipótesis nula es verdadera, el valor de CMTR/CME parecerá ser un valor tomado de esta distribución F.

Pero, si la hipótesis nula es falsa, el valor de CMTR/CME será muy grande debido a que CMTR sobreestima σ^2. Por tanto, si el valor de CMTR/CME resulta ser demasiado grande para haber sido tomado de la distribución F con $k - 1$ grados de libertad en el numerador y $n_T - k$ grados de libertad en el denominador, se rechazará H_0. Como la decisión de rechazar H_0 está basada en el valor de CMTR/CME, el estadístico de prueba que se usa para probar la igualdad de k poblaciones es el siguiente.

ESTADÍSTICO DE PRUEBA PARA LA IGUALDAD DE *k* MEDIAS POBLACIONALES

$$F = \frac{\text{CMTR}}{\text{CME}} \tag{13.12}$$

Este estadístico de prueba sigue una distribución F con $k - 1$ grados de libertad en el numerador y $n_T - k$ grados de libertad en el denominador.

Ahora recuerde el experimento de Chemitech, se usará $\alpha = 0.05$ para realizar la prueba de hipótesis. El valor del estadístico de prueba es

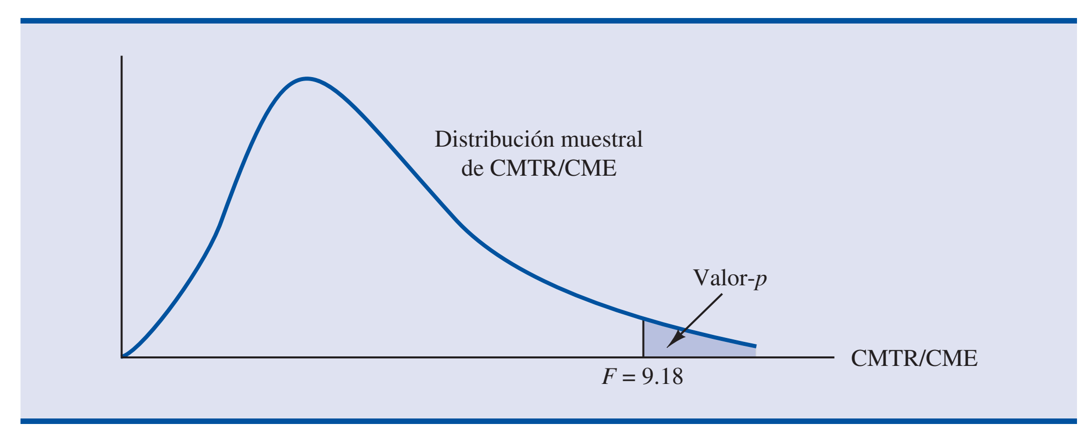
$$F = \frac{\text{CMTR}}{\text{CME}} = \frac{260}{28.33} = 9.18$$

Los grados de libertad para el numerador son $k - 1 = 3 - 1 = 2$ y los grados de libertad para el denominador son $n_T - k = 15 - 3 = 12$. Como la hipótesis nula sólo se rechazará si se obtiene un valor grande para el estadístico de prueba, el valor-p será el área en la cola superior de la distribución F a la derecha del estadístico de prueba $F = 9.18$. En la figura 13.4 se muestra la distribución muestral de F = CMTR/CME, el valor del estadístico de prueba y el área en la cola superior que es el valor-p de esta prueba de hipótesis.

Área en la cola superior	**0.10**	**0.05**	**0.025**	**0.01**
Valor F ($gl_1 = 2$, $gl_2 = 12$)	2.81	3.89	5.10	6.93

$F = 9.18$

FIGURA 13.4 CÁLCULO DEL VALOR-*p* A PARTIR DE LA DISTRIBUCIÓN MUESTRAL DE CMTR/CME



En el apéndice F se muestra cómo calcular el valor-p empleando Excel o Minitab.

En la tabla 4 del apéndice B se encuentran las áreas siguientes en la cola superior de la distribución F con 2 grados de libertad en el numerador y 12 grados de libertad en el denominador. Como $F = 9.18$ es mayor que 6.93, el área en la cola superior, correspondiente a $F = 9.18$ es menor que 0.01. Por tanto, el valor-*p* es menor que 0.01. Para obtener el valor-*p* exacto, que es 0.004, se puede usar Minitab o Excel. Como el valor-$p \leq \alpha = .05$, se rechaza H_0. La prueba proporciona evidencias suficientes para concluir que las medias de las tres poblaciones no son iguales. En otras palabras, el análisis de varianza favorece la conclusión de que las medias poblacionales del número de unidades producidas por semana, con cada uno de los tres métodos para armar los sistemas de filtración, no son iguales.

Como se hace en otras pruebas de hipótesis, también puede emplearse aquí el método del valor crítico. Como $\alpha = 0.05$, el valor crítico de F es el que deja un área de 0.05 en la cola superior de la distribución F con 2 y 12 grados de libertad. En las tablas de la distribución F se encuentra $F_{0.05} = 3.89$. Por tanto, la regla de rechazo en el caso del experimento de Chemitech es

$$\text{Rechazar } H_0 \text{ si } F \geq 3.89$$

Como $F = 9.18$, se rechaza H_0 y se concluye que las medias de las tres poblaciones no son iguales. A continuación se presenta un resumen del procedimiento para probar la igualdad de k medias poblacionales.

PRUEBA DE LA IGUALDAD DE k MEDIAS POBLACIONALES

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_k$$
H_a: No todas las medias poblacionales son iguales

ESTADÍSTICO DE PRUEBA

$$F = \frac{\text{CMTR}}{\text{CME}}$$

REGLA DE RECHAZO

Método del valor-p: Rechazar H_0 si el valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $F \geq F_\alpha$

donde el valor de F_α está basado en una distribución F con $k - 1$ grados de libertad en el numerador y $n_T - k$ grados de libertad en el denominador.

Tabla de ANOVA

Para presentar de manera adecuada los cálculos anteriores se usa una tabla conocida como tabla para el análisis de varianza o **tabla ANOVA**. En la tabla 13.2 se muestra la forma general de una tabla ANOVA para un diseño completamente aleatorizado. En la tabla 13.3 se presenta la tabla ANOVA correspondiente al experimento de Chemitech. A la suma de los cuadrados de la fuente de variación que se indica como "Total" se le conoce como suma de cuadrados del total (STC). Observe que los resultados del experimento de Chemitech indican que STC = SCTR + SCE y que los grados de libertad que corresponden a esta suma total de cuadrados es la suma de los grados de libertad correspondientes a la suma de cuadrados debidos a los tratamientos más la suma de cuadrados debidas al error.

Cabe hacer notar que la STC dividida entre los grados de libertad $n_T - 1$ no es otra cosa que la varianza muestral general que se obtendría si se considerara la muestra de las 15 observaciones como un solo conjunto de datos. Si se considera todo el conjunto de datos como una sola muestra, la fórmula para calcular la suma de cuadrados del total, STC, es

$$\text{STC} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\bar{x}})^2 \tag{13.13}$$

Se puede demostrar que estos resultados observados para el análisis de varianza en el caso del experimento de Chemitech aplican también a otros problemas. Es decir,

$$\text{STC} = \text{SCTR} + \text{SCE} \tag{13.14}$$

El análisis de varianza puede entenderse como un procedimiento estadístico de partición de la suma total de los cuadrados en componentes separados.

En otras palabras, STC se parte en dos sumas de cuadrados: la suma de cuadrados debidas a los tratamientos y la suma de cuadrados debidas al error. Observe, además, que los grados de libertad que corresponden a la STC, $n_T - 1$, se pueden partir en grados de libertad correspondientes a SCTR, $k - 1$ y en grados de libertad correspondientes a SCE, $n_T - k$. El análisis de varianza se puede ver como el proceso de **partición de** la suma total de cuadrados y los grados de libertad en sus fuentes correspondientes: tratamiento y error. Al dividir las sumas de cuadrados entre los correspondientes grados de libertad, se obtienen las estimaciones de la varianza, el valor de F y el valor-p empleados en la prueba de hipótesis de igualdad entre las medias poblacionales.

TABLA 13.2 TABLA ANOVA PARA UN DISEÑO COMPLETAMENTE ALEATORIZADO

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Tratamientos	SCTR	$k - 1$	$\text{CMTR} = \dfrac{\text{SCTR}}{k-1}$	$\dfrac{\text{CMTR}}{\text{CME}}$	
Error	SCE	$n_T - k$	$\text{CME} = \dfrac{\text{SCE}}{n_T - k}$		
Total	STC	$n_T - 1$			

TABLA 13.3 TABLA DE ANÁLISIS DE VARIANZA PARA EL EXPERIMENTO DE CHEMITECH

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Tratamientos	520	2	260.00	9.18	0.004
Error	340	12	28.33		
Total	860	14			

FIGURA 13.5 SALIDA DE MINITAB PARA EL ANÁLISIS DE VARIANZA DEL EXPERIMENTO DE CHEMITECH

```
Source     DF       SS       MS       F       P
Factor      2    520.0    260.0    9.18   0.004
Error      12    340.0     28.3
Total      14    860.0

S = 5.323    R-Sq = 60.47%    R-Sq(adj) = 53.88%

                                 Individual 95% CIs For Mean Based on
                                 Pooled StDev
 Level      N     Mean    StDev  ---+---------+---------+---------+------
 A          5   62.000    5.244               (-------*-------)
 B          5   66.000    4.148                      (------*-------)
 C          5   52.000    5.568  (------*-------)
                                 ---+---------+---------+---------+------
Pooled StDev = 5.323              49.0      56.0      63.0      70.0
```

Resultados de computadora para el análisis de varianza

Cuando se tienen muestras grandes o una cantidad grande de poblaciones, los cálculos del análisis de varianza se realizan con más facilidad mediante paquetes de software para estadística. En los apéndices 13.1 y 13.2 se indican los pasos necesarios para los cálculos del análisis de varianza con Minitab o Excel. En la figura 13.5, aplicado al experimento de Chemitech, se presenta la pantalla de resultados de Minitab. En la primera parte de la pantalla se observa el formato ya conocido de la tabla de ANOVA. Si se compara la figura 13.5 con la tabla 13.3, se ve que la información es la misma, aunque algunos de los encabezados son ligeramente diferentes. *Source* se usa como encabezado de la columna correspondiente a Fuentes de variación; *Factor* corresponde al renglón Tratamientos, y las columnas correspondientes a grados de libertad y a las sumas de cuadrados están intercambiadas.

Observe que abajo de la tabla de ANOVA Minitab proporciona los respectivos tamaños de las muestras, las medias muestrales y las desviaciones estándar muestrales. Además proporciona una figura con la estimación por intervalos de 95% de confianza para cada una de las medias poblacionales. Para obtener la estimación de estos intervalos, Minitab emplea el CME como estimación de σ^2. Por tanto, la raíz cuadrada del CME proporciona la mejor estimación de la desviación estándar poblacional σ. En la pantalla de los resultados de Minitab, esta estimación de σ es la Pooled StDev; su valor es 5.323. Para mostrar cómo se calcula la estimación por intervalos se hará aquí la estimación por intervalo de 95% de confianza para la media poblacional del método A.

De acuerdo con lo visto en el estudio de intervalos de confianza en el capítulo 8, se sabe que la forma general de una estimación por intervalo para una media poblacional es

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \quad \textbf{(13.15)}$$

donde s es la estimación de la desviación estándar poblacional σ. Como la mejor estimación de σ es la dada por la Pooled StDev, se usa 5.323 en la expresión (13.15) como valor de s. Los grados de libertad para el valor de t son 12, los grados de libertad correspondientes a la suma de los cuadrados del error. Por tanto, como $t_{0.025} = 2.179$ se obtiene

$$62 \pm 2.179 \frac{5.323}{\sqrt{5}} = 62 \pm 5.19$$

Así, el intervalo de 95% de confianza para el método A va de $62 - 5.19 = 56.81$ a $62 + 5.19 = 67.19$. Como en el experimento de Chemitech los tamaños muestrales son iguales, también los intervalos de confianza para los métodos B y C se obtienen al sumar y restar 5.19 de la respectiva media muestral. En la salida de Minitab se aprecia que los anchos de los intervalos de confianza son los mismos.

Prueba para la igualdad de *k* medias poblacionales: un estudio observacional

Se ha visto el uso del análisis de varianza para probar la igualdad de *k* medias poblacionales cuando se emplea un diseño experimental completamente aleatorizado. Es importante notar que el ANOVA también se puede usar para probar la igualdad de tres o más medias poblacionales usando datos de un estudio observacional. Para dar un ejemplo, se considerará el caso de National Computer Products, Inc. (NCP.)

NCP fabrica, en sus tres fábricas situadas en Atlanta, Dallas y Seattle, impresoras y faxes. Con el fin de medir los conocimientos que tienen los empleados de estas tres plantas acerca de la administración de la calidad, se toma una muestra aleatoria de seis empleados de cada fábrica y se les aplica un examen acerca de su conocimiento de la calidad. En la tabla 13.4 se presentan las puntuaciones obtenidas en los exámenes por los 18 empleados. En esta tabla se dan también la media, la varianza y la desviación estándar muestrales de cada grupo. Los administradores de la empresa quieren usar estos datos para probar la hipótesis de que la media de las puntuaciones de los exámenes es la misma en las tres fábricas.

Como población 1 se define a los empleados de Atlanta, como población 2 a los de Dallas y como población 3 a los de Seattle. Sean

$$\mu_1 = \text{media de las puntuaciones de la población 1}$$
$$\mu_2 = \text{media de las puntuaciones de la población 2}$$
$$\mu_3 = \text{media de las puntuaciones de la población 3}$$

Aunque los verdaderos valores de μ_1, μ_2 y μ_3, nunca puedan conocerse, se usarán los resultados muestrales para probar las hipótesis siguientes.

$$H_0: \mu_1 = \mu_2 = \mu_3$$
$$H_a: \text{No todas las medias poblacionales son iguales}$$

Observe que la prueba de hipótesis para el estudio observacional de NCP es exactamente igual a la prueba de hipótesis para el experimento de Chemitech. También, para analizar los datos del es-



NCP

TABLA 13.4 PUNTUACIONES EN LOS EXÁMENES DE 18 EMPLEADOS

	Fábrica 1 **Atlanta**	**Fábrica 2** **Dallas**	**Fábrica 3** **Seattle**
	85	71	59
	75	75	64
	82	73	62
	76	74	69
	71	69	75
	85	82	67
Media muestral	79	74	66
Varianza muestral	34	20	32
Desviación estándar muestral	5.83	4.47	5.66

En el ejercicio 8 se le pide al lector que use el análisis de varianza para analizar los datos de NCP.

tudio observacional se emplea la misma metodología de análisis de varianza usada para analizar el experimento de Chemitech.

Aun cuando en ambos casos se usa la misma metodología del ANOVA, vale la pena observar la diferencia entre el estudio estadístico observacional de NCP y el estudio estadístico experimental de Chemitech. Las personas que realizaron el estudio de NCP no tuvieron control sobre la asignación de las fábricas a cada uno de los empleados. Las plantas ya funcionaban y cada uno de los empleados trabajaba en una de las tres fábricas. Lo único que se pudo hacer en este caso, fue tomar una muestra aleatoria de seis empleados de cada una de las fábricas y aplicarles el examen sobre conocimiento de la calidad. Para poder clasificarlo como un estudio experimental, NPC tendría que haber tomado al azar 18 empleados y después, de manera aleatoria, asignar las fábricas a cada empleado.

NOTAS Y COMENTARIOS

1. La media muestral general también se calcula como media ponderada de las k medias muestrales.

$$\bar{\bar{x}} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_T}$$

En los problemas en que se proporcionan las medias muestrales, para calcular la media general, es más sencillo usar esta fórmula que la expresión (13.3).

2. Si todas las muestras constan de n observaciones, la ecuación (13.6) se escribe como

$$\text{CMTR} = \frac{n\sum_{j=1}^{k}(\bar{x}_j - \bar{\bar{x}})^2}{k-1} = n\left[\frac{\sum_{j=1}^{k}(\bar{x}_j - \bar{\bar{x}})^2}{k-1}\right] = ns_{\bar{x}}^2$$

Observe que este resultado es el mismo que el presentado en la sección 13.1 cuando se presentó el concepto de estimación de σ^2 entre tratamientos. La ecuación (13.6) sólo es una generalización de este resultado para el caso de los tamaños muestrales distintos.

3. Si cada muestra tiene n observaciones, $n_T = kn$; por tanto, $n_T - k = k(n-1)$, y la ecuación (13.9) se expresa como

$$\text{CME} = \frac{\sum_{j=1}^{k}(n-1)s_j^2}{k(n-1)} = \frac{(n-1)\sum_{j=1}^{k}s_j^2}{k(n-1)} = \frac{\sum_{j=1}^{k}s_j^2}{k}$$

En otras palabras, si los tamaños muestrales son iguales, el CME es simplemente el promedio de las k varianzas muestrales. Observe que éste es el mismo resultado que se usó en la sección 13.1 cuando se presentó el concepto de estimación de σ^2 dentro de los tratamientos.

Ejercicios

Métodos



1. Los datos siguientes se obtuvieron de un diseño completamente aleatorizado.

	Tratamiento A	Tratamiento B	Tratamiento C
	162	142	126
	142	156	122
	165	124	138
	145	142	140
	148	136	150
	174	152	128
Media muestral	156	142	134
Varianza muestral	164.4	131.2	110.4

a. Calcule la suma de cuadrados entre tratamientos.
b. Calcule el cuadrado medio entre tratamientos.

c. Calcule la suma de cuadrados debida al error.
d. Calcule el cuadrado medio debido al error.
e. Dé la tabla de ANOVA para este problema.
f. Con $\alpha = 0.05$ pruebe si las medias de los tres tratamientos son iguales.

2. En un diseño completamente aleatorizado, para cada uno de los cinco niveles del factor se usaron siete unidades experimentales. Complete la tabla ANOVA siguiente.

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	***F***	**Valor-*p***
Tratamientos	300				
Error					
Total	460				

3. Vuelva al ejercicio 2.
   a. ¿Cuáles son las hipótesis en este problema?
   b. Utilice el nivel de significancia $\alpha = 0.05$, ¿ la hipótesis nula del inciso a se puede rechazar? Explique.

4. En un experimento diseñado para probar los niveles de tres tratamientos diferentes, se obtuvieron los resultados siguientes: STC = 400, SCTR = 150, $n_T = 19$. Dé la tabla ANOVA y pruebe si hay alguna diferencia significativa entre las medias de los resultados de los tres tratamientos. Use $\alpha = 0.05$.

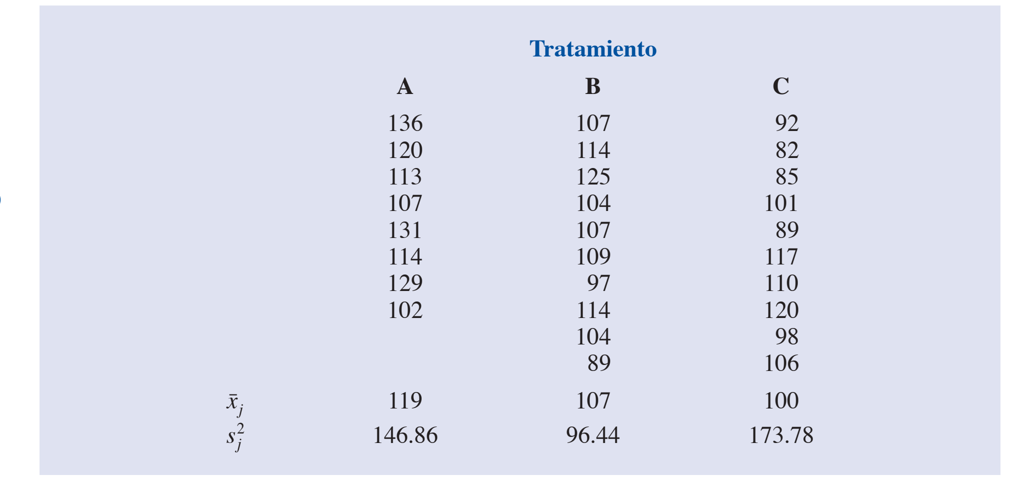
5. En un diseño completamente aleatorizado se usaron 12 unidades experimentales para el primer tratamiento, 15 para el segundo y 20 para el tercero. Complete el análisis de varianza siguiente. Emplee 0.05 como nivel de significancia, ¿hay diferencia significativa entre los tres tratamientos?

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	***F***	**Valor-*p***
Tratamientos	1200				
Error					
Total	1800				

6. Realice los cálculos del análisis de varianza para el siguiente diseño completamente aleatorizado. Con $\alpha = 0.05$, ¿ la diferencia entre las medias de tratamiento es significativa?



Exer6



	Tratamiento		
	A	**B**	**C**
	136	107	92
	120	114	82
	113	125	85
	107	104	101
	131	107	89
	114	109	117
	129	97	110
	102	114	120
		104	98
		89	106
$\bar{x}_j$	119	107	100
s_j^2	146.86	96.44	173.78

Aplicaciones

7. Un ingeniero propone tres métodos distintos para fabricar un producto. Para determinar el número de unidades producidas correctamente con cada método, se seleccionan al azar 30 empleados y se asignan de manera aleatoria a los tres métodos propuestos, de manera que cada método sea empleado por 10 trabajadores. Se anota el número de unidades producidas correctamente y a estos datos se aplica el análisis de varianza. Los resultados son los siguientes: STC = 10 800; SCTR = 4 560.
   a. Dé la tabla ANOVA de este problema.
   b. Use $\alpha = 0.05$ para determinar si existen diferencias significativas entre las medias de los tres métodos de fabricación
8. Vaya a la tabla 13.4 que presenta los datos de la NCP. Dé la tabla ANOVA y pruebe si existe diferencia significativa entre las medias de las puntuaciones de examen de las tres fábricas. Use $\alpha = 0.05$.
9. Para estudiar el efecto de la temperatura en el rendimiento de un proceso químico, se produjeron cinco lotes con cada uno de los tres tratamientos. Los resultados se presentan a continuación. Dé la tabla para el análisis de varianza. Use $\alpha = 0.05$ para probar si la temperatura afecta el rendimiento medio del proceso.

	Temperatura	
50°C	**60°C**	**70°C**
34	30	23
24	31	28
36	34	28
39	23	30
32	27	31

10. En una auditoría los auditores tienen que dar opiniones acerca de diversos aspectos con base en sus propias experiencias directas, indirectas o en una combinación de ambas. En un estudio se pidió a auditores que dieran su opinión acerca de la frecuencia con que se presentan errores en una auditoría. Suponga que se obtuvieron los resultados que se presentan a continuación; valores bajos indican opiniones más acertadas.



Directa	Indirecta	Combinación
17.0	16.6	25.2
18.5	22.2	24.0
15.8	20.5	21.5
18.2	18.3	26.8
20.2	24.2	27.5
16.0	19.8	25.8
13.3	21.2	24.2

Use $\alpha = 0.05$ para determinar si el tipo de experiencia en que se basa la opinión afecta la calidad de la misma.

11. En la publicidad de tres pinturas se dice que tienen el mismo tiempo de secado. Para verificar esto, se prueban cinco muestras de cada una de las pinturas. Se registra el tiempo en minutos necesario para que el secado sea suficiente para la aplicación de una segunda mano. Los datos obtenidos son los siguientes.



Pintura 1	Pintura 2	Pintura 3	Pintura 4
128	144	133	150
137	133	143	142
135	142	137	135
124	146	136	140
141	130	131	153

Con $\alpha = 0.05$ como nivel de significancia, realice una prueba para determinar si la media de los tiempos de secado es la misma en todas las pinturas.

12. Una conocida revista de automovilismo tomó tres de los mejores automóviles medianos fabricados en Estados Unidos, los probó y los comparó en relación con varios criterios. En una prueba sobre rendimiento de la gasolina, se probaron cinco automóviles de cada marca en un recorrido de 500 millas, los datos de rendimiento, en millas por galón de gasolina, se presentan a continuación. Use $\alpha = 0.05$ para probar si la diferencia en el rendimiento medio, en millas por galón, entre los tres automóviles es significativa.

Automóviles		
A	B	C
19	19	24
21	20	26
20	22	23
19	21	25
21	23	27



Procedimiento de comparación múltiple

Cuando se emplea el análisis de varianza para probar si las medias de k poblaciones son iguales, rechazar la hipótesis nula sólo permite concluir que las medias poblacionales *no son iguales*. En algunos casos se necesita dar un paso más y determinar dónde están las diferencias. El propósito de esta sección es mostrar el uso de **procedimientos de comparación múltiple** para hacer comparaciones entre pares de medias poblacionales.

LSD de Fisher

Suponga que en un análisis de varianza se encuentran evidencias estadísticas para rechazar la hipótesis nula que plantea la igualdad de las medias poblacionales. En tal caso, para determinar dónde están las diferencias se puede emplear el procedimiento de la diferencia mínima significativa (LSD, por sus siglas en inglés) de Fisher. Con el fin de ilustrar el uso del procedimiento de la LSD de Fisher para la comparación de pares de medias poblacionales, se retoma el experimento de Chemitech, visto en la sección 13.1. A partir del análisis de varianza, se concluyó que el número medio de unidades producidas por semana no era el mismo con los tres métodos. En tal caso la siguiente pregunta es: se cree que hay diferencia entre los métodos pero, ¿dónde están las diferencias? Es decir, las medias que difieren ¿son las de las poblaciones 1 y 2? ¿O las de las poblaciones 1 y 3? ¿O las de las poblaciones 2 y 3?

En el capítulo 10 se presentó un procedimiento estadístico para probar la hipótesis de la igualdad de dos medias poblacionales. Con una pequeña modificación en la manera de evaluar

la varianza poblacional, el procedimiento de la LSD de Fisher que basa en el estadístico de prueba t presentado para el caso de dos poblaciones. En la tabla siguiente se resume el procedimiento de la LSD de Fisher.

PROCEDIMIENTO DE LA LSD DE FISHER

$$H_0\colon \mu_i = \mu_j$$
$$H_a\colon \mu_i \neq \mu_j$$

ESTADÍSTICO DE PRUEBA

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\text{CME}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{13.16}$$

REGLA DE RECHAZO

Método del valor-p:	Rechazar H_0 si el valor-$p \leq \alpha$
Método del valor crítico:	Rechazar H_0 si $t \leq -t_{\alpha/2}$ o $t \geq t_{\alpha/2}$

donde el valor $t_{\alpha/2}$ se basa en la distribución t con $n_T - k$ grados de libertad.

A continuación se usará este procedimiento para determinar si existe alguna diferencia significativa entre la media de la población 1 (método A) y la media de la población 2 (método B) con $\alpha = 0.05$ como nivel de significancia. En la tabla 13.1 se ve que la media obtenida con el método A es 62 y la media obtenida con el método B es 66. En la tabla 13.3 se presenta el valor del CME, que es 28.33; ésta es la estimación de σ^2 con 12 grados de libertad. Con los datos de Chemitech, el valor que se obtiene para el estadístico de prueba es

$$t = \frac{62 - 66}{\sqrt{28.33\left(\frac{1}{5} + \frac{1}{5}\right)}} = -1.19$$

Como se trata de una prueba de dos colas, el valor-p es el doble del área bajo la curva de la distribución t a la izquierda de $t = -1.19$. En la tabla 2 del apéndice B se encuentra la información siguiente para la distribución t con 12 grados de libertad.

Área en la cola superior	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
Valor t (12 gl)	0.873	1.356	1.782	2.179	2.681	3.055

$t = 1.19$

En la tabla de la distribución t sólo hay valores positivos de t. Sin embargo, como la distribución t es simétrica, se puede hallar el área bajo la curva a la derecha de $t = 1.19$ y duplicarla para hallar el valor-p que corresponde a $t = -1.19$. En esta tabla se ve que $t = 1.19$ se encuentra entre 0.20 y 0.10. Al duplicar estas cantidades, se tiene que el valor-p debe estar entre 0.40 y 0.20. Si emplea Minitab o Excel puede encontrar el valor-p exacto, que es 0.2571. Como el valor-p es mayor que $\alpha = 0.05$, no se puede rechazar la hipótesis nula. Por tanto, no se puede concluir que las medias poblacionales de los números de unidades producidas por semana con los métodos A y B sean diferentes.

En el apéndice F se muestra cómo calcular los valores-p usando Minitab o Excel.

Muchas personas encuentran más fácil determinar qué tan grande tiene que ser la diferencia entre las medias muestrales para que se rechace H_0. En este caso el estadístico de prueba es $\bar{x}_i - \bar{x}_j$, y la prueba se realiza siguiendo el procedimiento que se presenta a continuación.

PROCEDIMIENTO DE LA LSD DE FISHER BASADO EN EL ESTADÍSTICO DE PRUEBA $\bar{x}_i - \bar{x}_j$

$$H_0\colon \mu_i = \mu_j$$
$$H_a\colon \mu_i \neq \mu_j$$

ESTADÍSTICO DE PRUEBA

$$\bar{x}_i - \bar{x}_j$$

REGLA DE RECHAZO PARA EL NIVEL DE SIGNIFICANCIA α

$$\text{Rechazar } H_0 \text{ si } |\bar{x}_i - \bar{x}_j| \geq \text{LSD}$$

donde

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{CME}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{13.17}$$

En el experimento de Chemitech, el valor de la LSD es

$$\text{LSD} = 2.179\sqrt{28.33\left(\frac{1}{5} + \frac{1}{5}\right)} = 7.34$$

Observe que si todos los tamaños muestrales son iguales sólo se necesita calcular un valor de la LSD. En tales casos, basta comparar la magnitud de la diferencia entre dos medias muestrales con el valor de la LSD. Por ejemplo, la diferencia entre las medias muestrales de las poblaciones 1 (método A) y 3 (método C) es 62 – 52 = 10. Esta diferencia es mayor que la LSD = 7.34, lo que significa que se puede rechazar la hipótesis de que las medias del número de unidades producidas por semana con los métodos A y C sean iguales. De manera similar, entre las medias muestrales de las poblaciones 2 y 3, la diferencia es 66 – 52 = 14 > 7.34, y se puede rechazar la hipótesis de que las medias poblacionales obtenidas con el método B y con el método C sean iguales. Así, la conclusión es que tanto el método A como el método B difieren del método C.

La LSD de Fisher también se usa para obtener una estimación mediante un intervalo de confianza de la diferencia entre las medias de dos poblaciones. El procedimiento general que se emplea es el siguiente.

ESTIMACIÓN POR INTERVALO DE CONFIANZA DE LA DIFERENCIA ENTRE DOS MEDIAS POBLACIONES USANDO EL PROCEDIMIENTO DE LA LSD DE FISHER

$$\bar{x}_i - \bar{x}_j \pm \text{LSD} \tag{13.18}$$

donde

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{CME}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{13.19}$$

y $t_{\alpha/2}$ pertenece a la distribución t con $n_T - k$ grados de libertad.

Si el intervalo de confianza hallado con la expresión (13.18) incluye el valor cero, no se puede rechazar la hipótesis nula de que las dos medias poblacionales sean iguales. Pero, si el intervalo de confianza no incluye al valor cero, se puede concluir que sí hay diferencia entre las medias poblacionales. En el caso del experimento de Chemitech, recuerde que la LSD = 7.34 (que corresponde a $t_{0.025} = 2.179$). Por tanto, una estimación de la diferencia entre las medias poblacionales 1 y 2 empleando un intervalo de 95% de confianza es $62 - 66 \pm 7.34 = -4 \pm 7.34 = -11.34$ to 3.34; como este intervalo incluye el cero, no se puede rechazar la hipótesis de que las dos medias sean iguales.

Tasas de error tipo I

El estudio del procedimiento de la LSD de Fisher inició con la premisa de que el análisis de varianza proporcionaba evidencias estadísticas para rechazar la hipótesis nula de la igualdad entre medias poblacionales. Se mostró que en tales casos se puede emplear el procedimiento de la LSD de Fisher para determinar dónde están las diferencias. Técnicamente a este procedimiento se le conoce como prueba *restringida* o *protegida* de la LSD debido a que sólo se usa si primero se ha encontrado un valor F significativo al usar el análisis de varianza. Para ver porqué es importante esta distinción en las pruebas de comparación múltiple, es necesario explicar la diferencia entre tasa de error tipo I por *comparación* y tasa de error tipo I por *experimentación*.

En el experimento de Chemitech se usa el procedimiento de la LSD de Fisher para hacer tres pares de comparaciones

Prueba 1	**Prueba 2**	**Prueba 3**
$H_0\colon \mu_1 = \mu_2$	$H_0\colon \mu_1 = \mu_3$	$H_0\colon \mu_2 = \mu_3$
$H_a\colon \mu_1 \neq \mu_2$	$H_a\colon \mu_1 \neq \mu_3$	$H_a\colon \mu_2 \neq \mu_3$

En cada caso el nivel de significancia empleado es $\alpha = 0.05$. Por tanto, en cada prueba, si la hipótesis nula es verdadera, la probabilidad de que se cometa un error tipo I es $\alpha = 0.05$, entonces la probabilidad de no cometer un error tipo I es $1 - 0.05 = 0.95$. En el estudio de los procedimientos de comparación múltiple a esta probabilidad de cometer un error tipo I ($\alpha = 0.05$) se le conoce como **tasa de error tipo I por comparación**. Las tasas de error tipo I por comparación indican el nivel de significancia que corresponde a una sola comparación por pares.

Considere ahora una cuestión ligeramente diferente. ¿Cuál es la probabilidad de que al hacer tres comparaciones por pares se cometa un error tipo I en por lo menos una de las tres pruebas? Para responder esta pregunta, observe que la probabilidad de que no se cometa un error tipo I en ninguna de las tres pruebas es (0.95)(0.95)(0.95) = 0.8574.* Por tanto, la probabilidad de cometer por lo menos un error tipo I es $1 - 0.8574 = 0.1426$. Entonces, cuando se usa el procedimiento de la LSD de Fisher para hacer los tres pares de comparaciones, la tasa de error tipo I correspondiente a este método no es 0.05, sino 0.1426; a esta tasa de error se le conoce como **tasa de error tipo I por experimentación** o *general*. Para evitar confusiones, la tasa de error tipo I por experimentación se denota α_{EW}.

La tasa de error tipo I por experimentación es mayor en problemas con más poblaciones. Por ejemplo, en un problema con cinco poblaciones hay 10 pares de comparaciones. Si se prueban todas las comparaciones posibles por pares usando el procedimiento de la LSD de Fisher con una tasa de error por comparación de $\alpha = 0.05$, la tasa de error tipo I por experimentación será $1 - (1 - 0.05)^{10} = 0.40$. En tales casos se prefiere buscar otras alternativas que proporcionen un mejor control sobre la tasa de error por experimentación.

Una alternativa para controlar la tasa de error general por experimentación, conocida como el ajuste de Bonferroni, consiste en usar en cada prueba tasas de error por comparación más pe-

*Se supone que las tres pruebas son independientes y, por tanto, la probabilidad conjunta de los tres eventos se obtiene con la simple multiplicación de las probabilidades individuales. De hecho, las tres pruebas no son independientes porque CME se usa en cada prueba; en consecuencia, el error supuesto es mayor que el error mostrado.

queñas. Por ejemplo, si se quieren probar C comparaciones por pares y se desea que la probabilidad máxima de cometer un error tipo I en todo el experimento sea α_{EW} simplemente se usa una tasa de error por comparación igual a α_{EW}/C. En el experimento de Chemitech, si se desea emplear el procedimiento de la LSD de Fisher, para probar los tres pares de comparaciones con una tasa de error máximo por experimentación de $\alpha_{EW} = 0.05$, se establece como tasa de error por comparación $\alpha = 0.05/3 = 0.017$. En un problema con cinco poblaciones y 10 comparaciones por pares, el ajuste de Bonferroni sugeriría una tasa de error por comparación de $0.05/10 = 0.005$. Recuerde que cuando se estudiaron las pruebas de hipótesis en el capítulo 9 se vio que para un tamaño de muestra dado, toda disminución en la probabilidad de cometer un error tipo I aumenta la probabilidad de cometer un error tipo II, el cual corresponde a aceptar la hipótesis de que las dos medias poblacionales son iguales cuando en realidad no lo son. Por tanto suele haber una renuencia a realizar pruebas individuales con una baja tasa de error tipo I por comparación debido al aumento del riesgo de cometer un error tipo II.

Como solución para tales situaciones se han elaborado otros varios procedimientos como el procedimiento de Turkey y la prueba de rango múltiple de Duncan. Sin embargo, en la comunidad estadística existe una gran controversia respecto a cuál es el "mejor" procedimiento. La verdad es que no hay uno que sea el mejor para todo tipo de problemas.

Ejercicios

Métodos



13. Los datos siguientes se obtuvieron con un diseño completamente aleatorizado.

	Tratamiento A	**Tratamiento B**	**Tratamiento C**
	32	44	33
	30	43	36
	30	44	35
	26	46	36
	32	48	40
Media muestral	30	45	36
Varianza muestral	6.00	4.00	6.50

a. Con $\alpha = 0.05$ como nivel de significancia, ¿puede rechazar la hipótesis nula: las medias de los tres tratamientos son iguales?
b. Use el procedimiento LSD de Fisher para probar si existe una diferencia significativa entre las medias de los tratamientos A y B, A y C y B y C. Use $\alpha = 0.05$.
c. Use el procedimiento LSD de Fisher para obtener un intervalo de estimación de 95% de confianza para la diferencia entre las medias de los tratamientos A y B.

14. Los datos siguientes se obtuvieron con un diseño completamente aleatorizado. Para los cálculos siguientes use $\alpha = 0.05$.

	Tratamiento 1	**Tratamiento 2**	**Tratamiento 3**
	63	82	69
	47	72	54
	54	88	61
	40	66	48
$\bar{x}_j$	51	77	58
s_j^2	96.67	97.34	81.99

a. Use el análisis de varianza para probar si hay una diferencia significativa entre las medias de los tres tratamientos.
b. Use el procedimiento LSD de Fisher para probar cuáles son las medias que difieren.

Aplicaciones



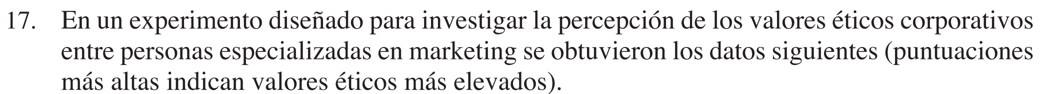
15. Con el fin de probar si la media del tiempo necesario para mezclar un lote de un material es la misma si emplea las máquinas de tres fabricantes, la empresa Jacobs Chemical obtiene los datos siguientes.

Fabricantes		
1	**2**	**3**
20	28	20
26	26	19
24	31	23
22	27	22

a. Utilice estos datos para probar si las medias de los tiempos necesarios para mezclar un lote de material usando las máquinas de estos tres fabricantes difieren. Use $\alpha = 0.05$.
b. Con $\alpha = 0.05$ como nivel de significancia, use el procedimiento LSD de Fisher para probar la igualdad entre las medias obtenidas con las máquinas del fabricante 1 y del fabricante 3. ¿Qué conclusión se obtiene después de realizar la prueba?




16. Vuelva al ejercicio 15, use el procedimiento LSD de Fisher para obtener una estimación por intervalo de 95% de confianza para la diferencia entre las medias de las poblaciones 1 y 2.


17. En un experimento diseñado para investigar la percepción de los valores éticos corporativos entre personas especializadas en marketing se obtuvieron los datos siguientes (puntuaciones más altas indican valores éticos más elevados).

Gerentes de marketing	Investigación en marketing	Publicidad
6	5	6
5	5	7
4	4	6
5	4	5
6	5	6
4	4	6

a. Use $\alpha = 0.05$ para probar si existe una diferencia significativa de percepción entre los tres grupos.
b. Con $\alpha = 0.05$, como nivel de significancia, se puede concluir que sí hay diferencias entre la percepción de los gerentes de marketing, los especialistas en investigación sobre marketing y los especialistas en publicidad. Emplee los procedimientos de esta sección para determinar dónde están las diferencias. Use $\alpha = 0.05$.

18. Para probar si existe una diferencia significativa entre cuatro máquinas respecto al número de horas entre dos averías se obtuvieron los datos siguientes.

Máquina 1	Máquina 2	Máquina 3	Máquina 4
6.4	8.7	11.1	9.9
7.8	7.4	10.3	12.8
5.3	9.4	9.7	12.1
7.4	10.1	10.3	10.8
8.4	9.2	9.2	11.3
7.3	9.8	8.8	11.5

a. Con $\alpha = 0.05$, como nivel de significancia, ¿cuál es la diferencia, si hay alguna, entre las medias poblacionales de los tiempos de las cuatro máquinas?
b. Use el procedimiento LSD de Fisher para probar la igualdad de las medias en las máquinas 2 y 4. Use 0.05 como nivel de significancia.

19. Vuelva al ejercicio 18. Use el ajuste de Bonferroni para probar si hay diferencia significativa entre todos los pares de muestras. Suponga que desea que el máximo de la tasa de error por experimentación sea 0.05.

20. *Condé Nast Traveler* realizó un sondeo anual en el que se pidió a los lectores que dieran una calificación para su crucero favorito. Se calificaron por separado embarcaciones pequeñas, (hasta 500 pasajeros), medianas (más de 500 pero menos de 1 500 pasajeros) y grandes (mínimo de 1 500 pasajeros). A continuación se presentan las calificaciones dadas a ocho embarcaciones de cada uno de los tamaños, chico, mediano y grande; las ocho embarcaciones de cada grupo fueron tomadas aleatoriamente. Todas las embarcaciones se calificaron con una escala de 100 en la que las puntuaciones más altas corresponden a mejor servicio (*Condé Nast Traveler*, febrero de 2003).



Ships

Embarcaciones pequeñas		Embarcaciones medianas		Embarcaciones grandes	
Nombre	**Calificación**	**Nombre**	**Calificación**	**Nombre**	**Calificación**
Hanseactic	90.5	Amsterdam	91.1	Century	89.2
Mississippi Queen	78.2	Crystal Symphony	98.9	Disney Wonder	90.2
Philae	92.3	Maasdam	94.2	Enchantment of the Seas	85.9
Royal Clipper	95.7	Noordam	84.3	Grand Princess	84.2
Seabourn Pride	94.1	Royal Princess	84.8	Infinity	90.2
Seabourn Spirit	100	Ryndam	89.2	Legend of the Seas	80.6
Silver Cloud	91.8	Statendam	86.4	Paradise	75.8
Silver Wind	95	Veendam	88.3	Sun Princess	82.3

a. Use $\alpha = 0.05$ para probar si existe alguna diferencia significativa entre las medias de las calificaciones dadas al servicio de cada uno de los grupos de tamaño.
b. Emplee los procedimientos de esta sección para determinar dónde están las diferencias. Use $\alpha = 0.05$.

13.4 Diseño de bloques aleatorizado

Hasta ahora sólo se ha considerado el diseño completamente aleatorizado. Como se recordará, para probar la diferencia entre las medias de los tratamientos se calcula el valor de F mediante el cociente

$$F = \frac{\text{CMTR}}{\text{CME}} \tag{13.20}$$

Un diseño completamente aleatorizado resulta útil cuando las unidades experimentales son homogéneas. Si las unidades experimentales son heterogéneas, se suele emplear ***la formación de bloques*** *para tener grupos homogéneos.*

Sin embargo, si diferencias debidas a factores extraños (factores no considerados en el experimento) hacen que en este cociente el término CME se vuelva más grande, pueden surgir problemas. En estos casos, en la ecuación (13.20) el valor de F será más pequeño, haciendo que se concluya que no hay diferencia entre las medias de los tratamientos cuando en realidad sí la hay.

En esta sección se presenta un diseño experimental conocido como **diseño de bloques aleatorizado**. El objetivo en este diseño es controlar algunas fuentes extrañas de variación eliminándolas del término CME. Este diseño tiende a proporcionar una mejor estimación de la varianza del error y conduce a pruebas de hipótesis más robustas en términos de la posibilidad de detec-

tar diferencias entre medias de tratamientos. Para ilustrar esto se verá un estudio sobre el estrés que experimentan los controladores del tráfico aéreo.

Prueba de estrés para los controladores del tráfico aéreo

Como resultado de un estudio para medir la fatiga y el estrés de los controladores del tráfico aéreo, se propusieron modificaciones y rediseños al puesto de trabajo. Después de considerar diversos diseños del puesto de trabajo, se seleccionaron tres alternativas consideradas con el mayor potencial para reducir el estrés en los controladores. La pregunta clave es: ¿En qué medida difieren estas tres alternativas en su efecto sobre el estrés de los controladores? Para responder esta pregunta se necesita diseñar un experimento que proporcione mediciones del estrés de los controladores del tráfico aéreo bajo cada una de estas alternativas.

En los estudios experimentales relacionados con negocios suelen intervenir unidades muy heterogéneas; en consecuencia los diseños de bloques aleatorizados se suelen emplear con frecuencia.

En el diseño experimental, la formación de bloques es similar a la estratificación en el muestreo.

Si se empleara un diseño completamente aleatorizado, una muestra aleatoria de controladores sería asignada a cada uno de los alternativos puestos de trabajo. Pero, se entiende que los controladores difieren sustancialmente en su habilidad para manejar situaciones estresantes. Lo que para un controlador es un gran estrés, para otro puede ser sólo un estrés moderado e incluso pequeño. Por tanto, al considerar la fuente de variación dentro del grupo (CME), hay que tener en cuenta que esta variación comprende tanto el error aleatorio como el error debido a las diferencias individuales de los controladores. En efecto, los administradores consideran que la variabilidad entre los controladores será la contribución principal al término CME.

Una manera de hacer a un lado el efecto de las diferencias individuales es usar el diseño de bloques aleatorizado. En ese diseño, se identifica la variabilidad debida a las diferencias individuales de los controladores y se elimina del término CME. En el diseño de bloques aleatorizado se emplea una sola muestra de controladores. Cada uno de los controladores de la muestra se prueba con cada una de las tres alternativas de puestos de trabajo. En la terminología del diseño de experimentos el puesto de trabajo es el *factor de interés* y los controladores son los *bloques*. Los tres tratamientos o poblaciones del factor puesto de trabajo son las tres alternativas de puesto de trabajo. Para simplificar, a las tres alternativas del puesto de trabajo se les designará como sistema A, sistema B y sistema C.

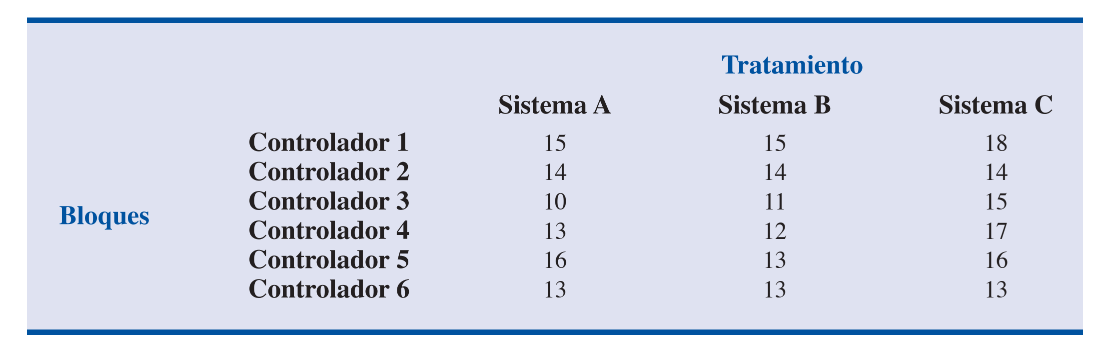
El aspecto *aleatorizado* en el diseño de bloques aleatorizado es el orden aleatorio en el que les son asignados los tratamientos (sistemas) a los controladores. Si cada controlador probara los tres sistemas en el mismo orden, cualquier diferencia encontrada entre los sistemas podría deberse al orden de la prueba y no a las verdaderas diferencias entre los sistemas.

Para obtener los datos necesarios, en el Centro de Control Cleveland en Oberlin, Ohio, se instalaron las tres alternativas de puesto de trabajo. Se seleccionaron seis controladores en forma aleatoria y se asignó cada uno a uno de los sistemas para que lo operara. Después de practicar una entrevista y un examen médico a cada uno de los participantes en el estudio se obtuvieron las mediciones del estrés de cada controlador en cada uno de los sistemas. En la tabla 13.5 se presentan estos datos.

En la tabla 13.6 aparece un resumen de los datos de estrés recolectados. En esta tabla se presentan también los totales de las columnas (tratamientos) y los totales de los renglones (bloques)

TABLA 13.5 DISEÑO DE BLOQUES ALEATORIZADO PARA LA PRUEBA DE ESTRÉS EN LOS CONTROLADORES DE TRÁFICO AÉREO.





		Tratamiento		
		Sistema A	**Sistema B**	**Sistema C**
Bloques	**Controlador 1**	15	15	18
	Controlador 2	14	14	14
	Controlador 3	10	11	15
	Controlador 4	13	12	17
	Controlador 5	16	13	16
	Controlador 6	13	13	13

TABLA 13.6 RESUMEN DE LOS DATOS DE ESTRÉS OBTENIDOS EN LA PRUEBA DE ESTRÉS APLICADA A LOS CONTROLADORES AÉREOS

		Tratamientos			Total del renglón o del bloque	Media del bloque
		Sistema A	Sistema B	Sistema C		
Bloques	Controlador 1	15	15	18	48	$\bar{x}_{1\cdot} = 48/3 = 16.0$
	Controlador 2	14	14	14	42	$\bar{x}_{2\cdot} = 42/3 = 14.0$
	Controlador 3	10	11	15	36	$\bar{x}_{3\cdot} = 36/3 = 12.0$
	Controlador 4	13	12	17	42	$\bar{x}_{4\cdot} = 42/3 = 14.0$
	Controlador 5	16	13	16	45	$\bar{x}_{5\cdot} = 45/3 = 15.0$
	Controlador 6	13	13	13	39	$\bar{x}_{6\cdot} = 39/3 = 13.0$
Total de la columna o tratamiento		81	78	93	252	$\bar{\bar{x}} = \frac{252}{18} = 14.0$
Media del tratamiento		$\bar{x}_{\cdot 1} = \frac{81}{6} = 13.5$	$\bar{x}_{\cdot 2} = \frac{78}{6} = 13.0$	$\bar{x}_{\cdot 3} = \frac{93}{6} = 15.5$		

así como algunas medias muestrales necesarias que serán útiles para hacer los cálculos de la suma de los cuadrados del ANOVA. Como valores bajos de estrés se consideran mejores, los datos muestrales parecen favorecer al sistema B, en el que la media de las mediciones del estrés es 13. Sin embargo, la pregunta persiste: ¿los resultados muestrales justifican la conclusión de que las medias poblacionales de los niveles de estrés, con estos tres sistemas, difieren? Es decir, ¿son las diferencias estadísticamente significativas? Para responder esta pregunta estadística se emplea un análisis del cálculo de la varianza, similar al empleado en el diseño completamente aleatorizado.

Procedimiento ANOVA

En el procedimiento ANOVA para el diseño de bloques aleatorizados se requiere que se haga una partición de la suma total de cuadrados (STC) en tres grupos: suma de cuadrados debidas a los tratamientos, suma de cuadrados debidas a los bloques y suma de cuadrado debidas al error. La fórmula para esta partición es la siguiente.

$$\text{STC} = \text{SCTR} + \text{SCBL} + \text{SCE} \tag{13.21}$$

Esta suma de la partición de cuadrados se presenta en la tabla ANOVA para el diseño de bloques aleatorizado, como se muestra en la tabla 13.7. La notación empleada en la tabla es

$$k = \text{número de tratamientos}$$
$$b = \text{número de bloques}$$
$$n_T = \text{tamaño muestral total } (n_T = kb)$$

Observe que en la tabla ANOVA también se muestra la partición de los $n_T - 1$ grados de libertad totales de manera que $k - 1$ grados de libertad correspondan a los tratamientos, $b - 1$ correspondan a los bloques y $(k - 1)(b - 1)$ correspondan al término del error. En la columna cuadrado medio se dan las sumas de los cuadrados divididas entre los grados de libertad y $F =$ CMTR/CME es el cociente F que se usa para probar si hay diferencias significativas entre las medias de los tratamientos. La contribución más importante del diseño de bloques aleatorizado es que, al emplear bloques se eliminan del término CME las diferencias individuales de los controladores y se obtiene una prueba más sólida para las diferencias de estrés entre las tres alternativas de puestos de trabajo.

TABLA 13.7 TABLA ANOVA PARA EL DISEÑO DE BLOQUES ALEATORIZADO CON k TRATAMIENTOS Y b BLOQUES

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Tratamientos	SCTR	$k-1$	$\text{CMTR} = \dfrac{\text{SCTR}}{k-1}$	$\dfrac{\text{CMTR}}{\text{CME}}$	
Bloques	SCBL	$b-1$	$\text{CMBL} = \dfrac{\text{SCBL}}{b-1}$		
Error	SCE	$(k-1)(b-1)$	$\text{CME} = \dfrac{\text{SCE}}{(k-1)(b-1)}$		
Total	STC	$n_T - 1$			

Cálculos y conclusiones

Para calcular el estadístico F que se necesita para probar si hay diferencia entre las medias de los tratamientos en un diseño de bloques aleatorizado, se necesita calcular el CMTR y el CME. Para calcular estos dos cuadrados medios, es necesario calcular primero SCTR y SCE; para esto también se calcula SCBL y STC. Para hacerlo más sencillo, estos cálculos se realizan en cuatro pasos. Además de la notación k, b y n_T ya introducida, se usará:

x_{ij} = valor de la observación correspondiente al tratamiento j en el bloque i.
$\bar{x}_{\cdot j}$ = media muestral con el tratamiento j
$\bar{x}_{i\cdot}$ = media muestral en el bloque i
$\bar{\bar{x}}$ = media muestral general

Paso 1. Calcular la suma total de cuadrados (STC)

$$\text{STC} = \sum_{i=1}^{b}\sum_{j=1}^{k}(x_{ij} - \bar{\bar{x}})^2 \tag{13.22}$$

Paso 2. Calcular la suma de los cuadrados debidos a los tratamientos (SCTR).

$$\text{SCTR} = b\sum_{j=1}^{k}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13.23}$$

Paso 3. Calcular la suma de los cuadrados debidos a los bloques (SCBL).

$$\text{SCBL} = k\sum_{i=1}^{b}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13.24}$$

Paso 4. Calcular la suma de cuadrados debidos al error (SCE).

$$\text{SCE} = \text{STC} - \text{SCTR} - \text{SCBL} \tag{13.25}$$

En el caso de los datos de la tabla 13.6 sobre los controladores del tráfico aéreo, con estos cálculos se obtienen las sumas de los cuadrados siguientes.

Paso 1. $\text{STC} = (15-14)^2 + (15-14)^2 + (18-14)^2 + \cdots + (13-14)^2 = 70$

Paso 2. $\text{SCTR} = 6[(13.5-14)^2 + (13.0-14)^2 + (15.5-14)^2] = 21$

Paso 3. $\text{SCBL} = 3[(16-14)^2 + (14-14)^2 + (12-14)^2 + (14-14)^2 + (15-14)^2 + (13-14)^2] = 30$

Paso 4. $\text{SCE} = 70 - 21 - 30 = 19$

TABLA 13.8 TABLA ANOVA PARA LA PRUEBA DEL ESTRÉS DE LOS CONTROLADORES DEL TRÁFICO AÉREO

Fuentes de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	Valor-*p*
Tratamientos	21	2	10.5	10.5/1.9 = 5.53	0.0241
Bloqueos	30	5	6.0		
Error	19	10	1.9		
Total	70	17			

Las sumas de cuadrados divididas entre sus grados de libertad dan los correspondientes cuadrados medios que se muestran en la tabla 13.8.

Ahora, para realizar la prueba de hipótesis, se va a usar $\alpha = 0.05$ como nivel de significancia. El valor del estadístico de prueba es

$$F = \frac{\text{CMTR}}{\text{CME}} = \frac{10.5}{1.9} = 5.53$$

Los grados de libertad en el numerador son $k - 1 = 3 - 1 = 2$ y los grados de libertad en el denominador son $(k - 1)(b - 1) = (3 - 1)(6 - 1) = 10$. Como la prueba de hipótesis se rechaza sólo cuando los valores del estadístico de prueba son grandes, el valor-*p* es el área bajo la distribución *F* a la derecha de $F = 5.53$. En la tabla 4 del apéndice B se encuentra que para 2 y 10 grados de libertad, $F = 5.53$ se encuentra entre $F_{0.025} = 5.46$ y $F_{0.01} = 7.56$. Por tanto, el área en la cola superior, o valor-*p,* se encuentra entre 0.01 y 0.025. Se pueden usar también Excel o Minitab y encontrar que el valor-*p* exacto para $F = 5.53$ es 0.0241. Como el valor-$p \leq \alpha = 0.05$, se rechaza la hipótesis nula H_0: $\mu_1 = \mu_2 = \mu_3$ y se concluye que las medias poblacionales de los niveles de estrés en las tres alternativas de puesto de trabajo no son iguales.

Acerca de este diseño de bloques aleatorizado se pueden hacer algunos comentarios generales. El diseño de experimentos descrito en esta sección es un diseño de bloques *completo*; la palabra “completo” indica que cada bloque se somete a todos los *k* tratamientos. Es decir, todos los controladores (bloques) fueron probados con los tres sistemas (tratamientos). A los diseños de experimentos en los que a cada bloque se le aplican algunos, pero no todos, los tratamientos se les llaman diseños de bloques *incompletos*. El estudio del diseño de bloques incompletos queda fuera del alcance de este libro.

Como en la prueba sobre el estrés de los controladores del tráfico aéreo, cada controlador usó todos los sistemas, este método garantiza un diseño de bloques completo. En algunos casos la formación de los bloques se realiza con unidades experimentales “similares” en cada bloque. Por ejemplo, suponga que en una prueba preliminar realizada a los controladores del tráfico aéreo, se divide la población de controladores en grupos que van desde personas con mucho estrés hasta individuos con estrés sumamente bajo. Aquí también se puede tener la formación de bloques haciendo que en el estudio participen tres controladores de cada nivel de estrés. En este caso, cada bloque consistirá en tres controladores de un mismo nivel de estrés. El aspecto aleatorizado del diseño debloques será la designación aleatoria de los tres controladores de cada bloque a los tres sistemas.

Por último, observe que en la tabla ANOVA que se presenta en la tabla 13.7, se da un valor *F* para probar los efectos de los tratamientos pero *no* de los bloques. La razón es que el experimento se diseñó para probar un solo factor: el diseño del puesto de trabajo. La formación de bloques basada en las diferencias del estrés individuales se hizo para eliminar tal variación del término CME. Pero, el estudio no se diseñó para detectar las diferencias individuales de estrés.

Algunos analistas calculan $F = \text{CMBL}/\text{CME}$ y usan este estadístico para probar la significancia de los bloques. Después usan los resultados como guía para determinar si el mismo tipo de bloques puede ser útil en experimentos futuros. Sin embargo, si la diferencia en el estrés de las personas ha de ser un factor en el estudio, deberá emplearse un diseño experimental diferente. Una prueba de significancia sobre los bloques no debe hacerse como una base para una conclusión acerca de un segundo factor.

NOTAS Y COMENTARIOS

En un diseño de bloques aleatorizado, los grados de libertad del error son menos que en un diseño completamente aleatorizado debido a que en los b bloques se pierden $b - 1$ grados de libertad. Si n es pequeño, los efectos potenciales debidos a los bloques pueden quedar ocultos debido a la pérdida de grados de libertad del error; con n grande los efectos se minimizan.

Ejercicios

Métodos



21. Considere los siguientes resultados experimentales obtenidos con un diseño de bloques aleatorizado. Realice los cálculos necesarios para dar la tabla de análisis de varianza.

		Tratamientos		
		A	**B**	**C**
Bloques	**1**	10	9	8
	2	12	6	5
	3	18	15	14
	4	20	18	18
	5	8	7	8

Use $\alpha = 0.05$ y realice la prueba para determinar si existe una diferencia significativa entre los tratamientos.

22. Los datos siguientes se obtuvieron mediante un diseño de bloques aleatorizado con cinco tratamientos y tres bloques: STC = 430, SCTR = 310, SCBL = 85. Dé la tabla ANOVA y realice una prueba para determinar si hay diferencia significativa entre los tratamientos.

23. Se realizó un experimento con cuatro tratamientos y ocho bloques. Complete la siguiente tabla de análisis de varianza.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadro medio	*F*
Tratamientos	900			
Bloques	400			
Error				
Total	1800			

Use $\alpha = 0.05$ y realice una prueba para determinar si existe diferencia significativa entre los tratamientos.

Aplicaciones

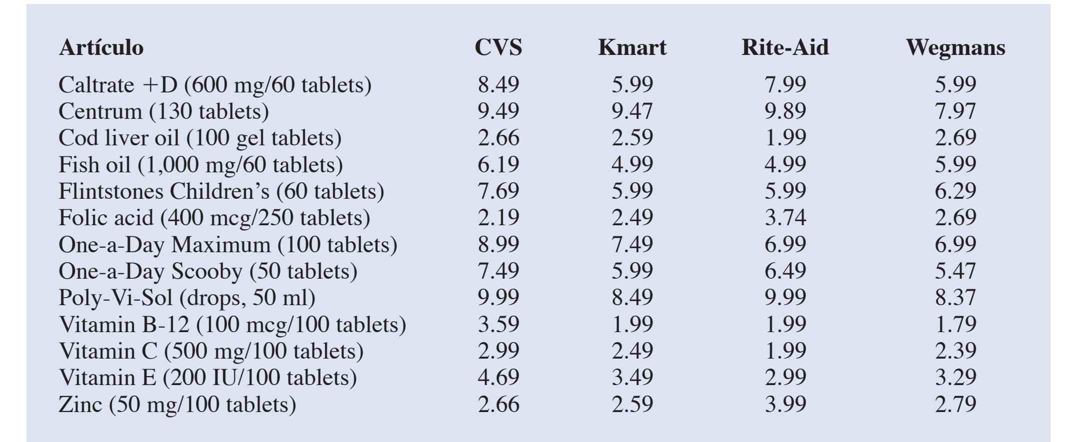
24. Un comerciante de automóviles realiza una prueba para determinar si la cantidad de tiempo en minutos que se necesita para una afinación de motor depende de si se emplea un analizador computarizado o un analizador electrónico. Los datos obtenidos son los siguientes.

		Analizador	
		Computarizado	**Electrónico**
Automóvil	**Compacto**	50	42
	Mediano	55	44
	Grande	63	46

Use $\alpha = 0.05$ y realice una prueba para determinar si existe diferencia significativa entre los tratamientos.

25. Durante los últimos años, los precios de las vitaminas y de otros complementos alimenticios aumentaron, también suele haber una gran variación de precios entre los distintos establecimientos. Los datos siguientes son los precios de 13 productos en cuatro establecimientos.





Artículo	**CVS**	**Kmart**	**Rite-Aid**	**Wegmans**
Caltrate +D (600 mg/60 tablets)	8.49	5.99	7.99	5.99
Centrum (130 tablets)	9.49	9.47	9.89	7.97
Cod liver oil (100 gel tablets)	2.66	2.59	1.99	2.69
Fish oil (1,000 mg/60 tablets)	6.19	4.99	4.99	5.99
Flintstones Children's (60 tablets)	7.69	5.99	5.99	6.29
Folic acid (400 mcg/250 tablets)	2.19	2.49	3.74	2.69
One-a-Day Maximum (100 tablets)	8.99	7.49	6.99	6.99
One-a-Day Scooby (50 tablets)	7.49	5.99	6.49	5.47
Poly-Vi-Sol (drops, 50 ml)	9.99	8.49	9.99	8.37
Vitamin B-12 (100 mcg/100 tablets)	3.59	1.99	1.99	1.79
Vitamin C (500 mg/100 tablets)	2.99	2.49	1.99	2.39
Vitamin E (200 IU/100 tablets)	4.69	3.49	2.99	3.29
Zinc (50 mg/100 tablets)	2.66	2.59	3.99	2.79

Use $\alpha = 0.05$ y realice una prueba para determinar si existe diferencia significativa entre los precios medios de los cuatro establecimientos.

26. Un factor importante en la elección de un procesador de palabras o de un sistema para la administración de datos es el tiempo necesario para aprender a usar el sistema. Para evaluar tres sistemas de administración de archivos, una empresa diseña una prueba con cinco operadores. Como se considera que la variabilidad entre los operadores es un factor significativo, se capacita a cada uno de los cinco operadores en cada uno de los tres sistemas de administración de archivos. A continuación se presentan los datos obtenidos

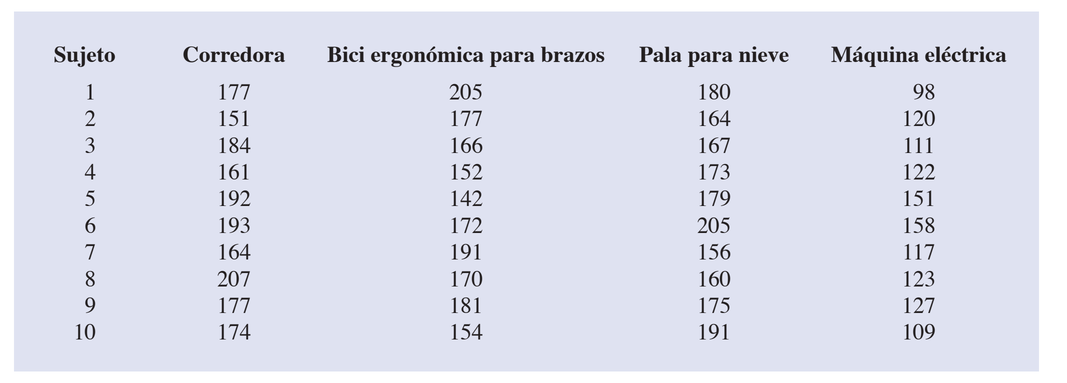
		Sistema		
		A	**B**	**C**
Operador	**1**	16	16	24
	2	19	17	22
	3	14	13	19
	4	13	12	18
	5	18	17	22

Use $\alpha = 0.05$ y realice una prueba para determinar si existe diferencia significativa entre los tiempos, en horas, necesarios para aprender a usar cada uno de los tres sistemas.

27. En un estudio publicado en el *Journal of the American Medical Association* se investigaba la demanda cardiaca al apalear grandes cantidades de nieve. Diez hombres saludables se sometieron a pruebas de ejercicio empleando una corredora y una bicicleta ergonómica para brazos. Después estos mismos hombres limpiaron dos tramos de nieve mojada y pesada con una pala ligera para nieve y una máquina eléctrica para despejar nieve. Se midió el ritmo cardiaco, la presión sanguínea y el consumo de oxígeno de cada uno de los participantes en la prueba, durante la remoción de nieve, y estos valores se compararon con los valores durante las pruebas de ejercicio. En la ta-

bla siguiente se presentan los valores de ritmo cardiaco, dados en pulsaciones por minuto, de cada uno de los 10 individuos.





Sujeto	Corredora	Bici ergonómica para brazos	Pala para nieve	Máquina eléctrica
1	177	205	180	98
2	151	177	164	120
3	184	166	167	111
4	161	152	173	122
5	192	142	179	151
6	193	172	205	158
7	164	191	156	117
8	207	170	160	123
9	177	181	175	127
10	174	154	191	109

Con $\alpha = 0.05$, como nivel de significancia, realice una prueba para determinar si existe diferencia significativa entre los diversos tratamientos.



Experimentos factoriales

Los diseños de experimentos vistos hasta ahora permiten obtener conclusiones estadísticas acerca de un solo factor. Sin embargo, en algunos experimentos se desean obtener conclusiones acerca de más de un factor o variable. Un **experimento factorial** es un diseño experimental que permite obtener, simultáneamente, conclusiones acerca de dos o más factores. El término *factorial* se emplea debido a que las condiciones experimentales comprenden todas las posibles combinaciones de los factores. Por ejemplo, si se tienen a niveles del factor A y b niveles del factor B, se obtendrán datos de ab combinaciones de tratamientos. En esta sección se verá un experimento factorial para dos factores. La idea básica se extiende a experimentos factoriales con más de dos factores.

Para ilustrar los experimentos factoriales para dos factores, se considerará un estudio realizado en relación con un examen de admisión para estudiantes con licenciatura que desean hacer un estudio sobre administración de negocios. Las puntuaciones que se pueden obtener en este examen de admisión van de 200 a 800, las puntuaciones más altas reflejan mejores aptitudes para el estudio en el futuro.

Como ayuda para la preparación de este examen, una institución ofrece los tres programas siguientes.

1. Una sesión de repaso de tres horas, en la que se revisa el tipo de preguntas que suelen encontrarse en el examen.
2. Un programa de un día en el que se ve el material más importante que se necesita saber para el examen y se hace un examen muestra que es incluso calificado.
3. Un curso intensivo de diez semanas en el que se determinan las debilidades de cada estudiante y se establece un programa individualizado para superar esas debilidades.

Por tanto, un factor en este estudio es el programa de preparación para el examen de admisión. Para este factor hay tres tratamientos: un repaso de tres horas, un programa de un día y un curso de 10 semanas. Se quiere determinar el efecto de cada uno de los programas sobre las puntuaciones obtenidas en este examen de admisión.

Por lo común este examen lo hacen estudiantes de tres licenciaturas, administración, ingeniería y ciencias. En consecuencia, el segundo factor que interesa en este estudio es si la licenciatura del estudiante influye en la calificación en el examen de admisión. Para este segundo factor hay también tres tratamientos, administración, ingeniería y ciencias. En el diseño factorial de este experimento en el que hay tres tratamientos para el factor A, programa de preparación, y tres tratamientos

TABLA 13.9 LAS NUEVE COMBINACIONES DE TRATAMIENTOS EN EL EXPERIMENTO CON DOS FACTORES DEL EXAMEN DE ADMISIÓN

		Factor B: licenciatura		
		Administración	**Ingeniería**	**Ciencias**
Factor A: programa de preparación	**Repaso de tres horas**	1	2	3
	Programa de un día	4	5	6
	Curso de 10 semanas	7	8	9

para el factor B, tipo de licenciatura, habrá en total $3 \times 3 = 9$ combinaciones de tratamientos. En la tabla 13.9 se resumen estas combinaciones de tratamientos o condiciones experimentales.

Suponga que se toma una muestra de dos estudiantes para cada una de las combinaciones de tratamientos de la tabla 13.9: dos estudiantes de administración participarán en el repaso de tres horas, dos participarán en el programa de un día y dos participarán en el curso de 10 semanas. Además, dos estudiantes de ingeniería y dos estudiantes de ciencias participarán en cada uno de los tres programas. En la terminología del diseño de experimentos, el tamaño muestral de dos para cada combinación de tratamientos indica que se tienen dos **replicaciones**. Se pueden usar también más replicaciones y tamaños muestrales mayores, pero en esta aplicación se quisieron minimizar los cálculos para hacer más claro este ejemplo.

En este diseño experimental se requiere que de *cada una* de las licenciaturas (administración, ingeniería y ciencias) se tomen aleatoriamente seis estudiantes que pretendan realizar este examen de admisión y, que después, dos estudiantes de cada licenciatura sean asignados de manera aleatoria a cada uno de los programas de preparación para el examen, con lo que en total participan 18 estudiantes en este estudio.

En la tabla 13.10 se presentan las puntuaciones obtenidas en el examen por estos estudiantes después de haber participado en los programas de preparación para el examen.

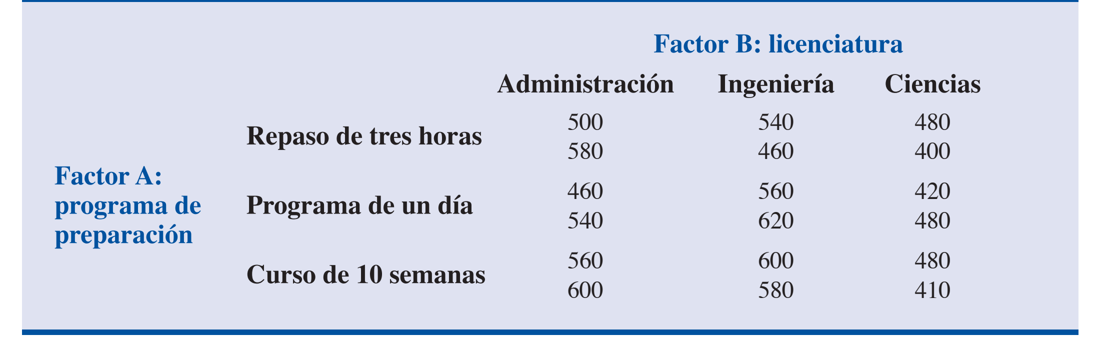
Los cálculos para el análisis de varianza con los datos de la tabla 13.10 permitirán responder las preguntas siguientes:

- **Efecto principal (factor A):** ¿Tienen los programas de preparación efectos diferentes sobre la puntuación obtenida en el examen de admisión?
- **Efecto principal (factor B):** ¿Tienen las licenciaturas efectos diferentes sobre la puntuación obtenida en el examen de admisión?
- **Efecto de interacción (factor A y B):** ¿Es uno de los programas de preparación mejor para los estudiantes que vienen de una de las tres licenciaturas, mientras que para los de otras licenciaturas es mejor otro de los programas?

El término **interacción** se refiere a un nuevo efecto que es posible estudiar debido a que se emplea un experimento factorial. Si el efecto interacción tiene algún impacto significante sobre

TABLA 13.10 PUNTUACIONES EN EL EXAMEN DE ADMISIÓN DEL EXPERIMENTO DE DOS FACTORES





		Factor B: licenciatura		
		Administración	**Ingeniería**	**Ciencias**
Factor A: programa de preparación	**Repaso de tres horas**	500	540	480
		580	460	400
	Programa de un día	460	560	420
		540	620	480
	Curso de 10 semanas	560	600	480
		600	580	410

TABLA 13.11 TABLA ANOVA PARA EL EXPERIMENTO FACTORIAL DE DOS FACTORES CON r REPLICACIONES

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Factor A	SCA	$a-1$	$\text{CMA} = \dfrac{\text{SCA}}{a-1}$	$\dfrac{\text{CMA}}{\text{CME}}$	
Factor B	SCB	$b-1$	$\text{CMB} = \dfrac{\text{SCB}}{b-1}$	$\dfrac{\text{CMB}}{\text{CME}}$	
Interacción	SCAB	$(a-1)(b-1)$	$\text{CMAB} = \dfrac{\text{SCAB}}{(a-1)(b-1)}$	$\dfrac{\text{CMAB}}{\text{CME}}$	
Error	SCE	$ab(r-1)$	$\text{CME} = \dfrac{\text{SCE}}{ab(r-1)}$		
Total	STC	$n_T - 1$			

las puntuaciones del examen de admisión, se podrá concluir que el efecto del tipo de programa de preparación depende de la licenciatura

Procedimiento ANOVA

El ANOVA para el experimento factorial de dos factores se parece al del experimento completamente aleatorizado y al del experimento con bloques aleatorizado en que también hay partición de las sumas de cuadrados y los grados de libertad en sus fuentes correspondientes. La fórmula para la partición de las sumas de cuadrados en sus diversos componentes se da a continuación.

$$\text{STC} = \text{SCA} + \text{SCB} + \text{SCAB} + \text{SCE} \tag{13.26}$$

En la tabla 13.11 se resumen las particiones de las sumas de cuadrados y de los grados de libertad. Se emplea la notación siguiente:

a = número de niveles (valores) del factor A
b = número de niveles (valores) del factor B
r = número de replicaciones
n_T = número total de observaciones realizadas en el experimento; $n_T = abr$

Cálculos y conclusiones

Para calcular los estadísticos F que se necesitan para las pruebas de significancia del factor A, del factor B y de la interacción, se necesitan calcular CMA, CMB, CMAB y CME. Para calcular estos cuatro cuadrados medios (o medias de cuadrados), se deben calcular primero SCA, SCB, SCAB y SCE; con esto se calcula también STC. Para simplificar, los cálculos se dividen en cinco pasos. Además de la notación ya introducida, se emplea la siguiente:

x_{ijk} = observación correspondiente a la réplica k del tratamiento i del factor A y del tratamiento j del factor B
$\bar{x}_{i\cdot}$ = media muestral de las observaciones del tratamiento i (factor A)
$\bar{x}_{\cdot j}$ = media muestral de las observaciones del tratamiento j (factor B)
$\bar{x}_{ij}$ = media muestral de las observaciones correspondientes a la combinación del tratamiento i (factor A) y del tratamiento j (factor B)
$\bar{\bar{x}}$ = media muestral general de todas las n_T observaciones

Paso 1. Calcular la suma total de cuadrados

$$\text{STC} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk} - \bar{\bar{x}})^2 \tag{13.27}$$

Paso 2. Calcular la suma de cuadrados del factor A

$$\text{SCA} = br\sum_{i=1}^{a}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13.28}$$

Paso 3. Calcular la suma de cuadrados del factor B

$$\text{SCB} = ar\sum_{j=1}^{b}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13.29}$$

Paso 4. Calcular la suma de cuadrados debida a la interacción

$$\text{SCAB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2 \tag{13.30}$$

Paso 5. Calcular la suma de cuadrados debida al error

$$\text{SCE} = \text{STC} - \text{SCA} - \text{SCB} - \text{SCAB} \tag{13.31}$$

En la tabla 13.12 se muestran los datos obtenidos en el experimento y las diversas sumas necesarias para los cálculos de las sumas de cuadrados. Mediante las ecuaciones (13.27) a (13.31), se calculan las diversas sumas de cuadrados del experimento factorial de dos factores del examen de admisión.

Paso 1. $\text{STC} = (500 - 515)^2 + (580 - 515)^2 + (540 - 515)^2 + \cdots + (410 - 515)^2 = 82\,450$

Paso 2. $\text{SCA} = (3)(2)[(493.33 - 515)^2 + (513.33 - 515)^2 + (538.33 - 515)^2] = 6100$

Paso 3. $\text{SCB} = (3)(2)[(540 - 515)^2 + (560 - 515)^2 + (445 - 515)^2] = 45\,300$

Paso 4. $\text{SCAB} = 2[(540 - 493.33 - 540 + 515)^2 + (500 - 493.33 - 560 + 515)^2] + \cdots + (445 - 538.33 - 445 + 515)^2] = 11\,200$

Paso 5. $\text{SCE} = 82\,450 - 6100 - 45\,300 - 11\,200 = 19\,850$

Estas sumas de cuadrados divididas entre sus grados de libertad dan los correspondientes cuadrados medios para estimar los dos efectos principales (programas de preparación y licenciatura) y el efecto de la interacción.

Para realizar la prueba de hipótesis de dos factores en el estudio del examen de admisión se usará el nivel de significancia, $\alpha = 0.05$. Debido a la gran cantidad de cálculos en un experimento factorial, la computadora tiene un papel importante en la realización de los cálculos necesarios en el análisis de varianza y para obtener los valores-p que se emplean para tomar las decisiones en la prueba de hipótesis. En la figura 13.6 se presenta la pantalla de resultados de Minitab para el análisis de varianza del experimento factorial de dos factores del examen de admisión. El valor-p para probar si hay diferencias significativas entre los tres programas de preparación (factor A) es 0.299. Como este valor-$p = 0.299$ es mayor a $\alpha = 0.05$, se concluye que los programas de preparación no hacen que haya diferencia significativa entre las medias de las puntuaciones obtenidas en los exámenes de admisión. Sin embargo, en relación con el efecto de la licenciatura, el valor-$p = 0.005$ es menor que $\alpha = 0.05$; por tanto, entre las tres licenciaturas

TABLA 13.12 RESUMEN DE LOS DATOS DEL EXAMEN DE ADMISIÓN PARA EL EXPERIMENTO DE DOS FACTORES

		Factor B: licenciatura				
	Totales de combinación de tratamiento	**Administración**	**Ingeniería**	**Ciencias**	**Totales de renglón**	**Medias del factor A**
Factor A: preparación del programa	**Repaso de tres horas**	500 580 1080 $\bar{x}_{11} = \frac{1080}{2} = 540$	540 460 1000 $\bar{x}_{12} = \frac{1000}{2} = 500$	480 400 880 $\bar{x}_{13} = \frac{880}{2} = 440$	2960	$\bar{x}_{1\cdot} = \frac{2960}{6} = 493.33$
	Programa de un día	460 540 1000 $\bar{x}_{21} = \frac{1000}{2} = 500$	560 620 1180 $\bar{x}_{22} = \frac{1180}{2} = 590$	420 480 900 $\bar{x}_{23} = \frac{900}{2} = 450$	3080	$\bar{x}_{2\cdot} = \frac{3080}{6} = 513.33$
	Curso de 10 semanas	560 600 1160 $\bar{x}_{31} = \frac{1160}{2} = 580$	600 580 1180 $\bar{x}_{32} = \frac{1180}{2} = 590$	480 410 890 $\bar{x}_{33} = \frac{890}{2} = 445$	3230	$\bar{x}_{3\cdot} = \frac{3230}{6} = 538.33$
	Totales de columna	3240	3360	2670	9270 ← Total general	
	Media del factor B	$\bar{x}_{\cdot 1} = \frac{3240}{6} = 540$	$\bar{x}_{\cdot 2} = \frac{3360}{6} = 560$	$\bar{x}_{\cdot 3} = \frac{2670}{6} = 445$	$\bar{\bar{x}} = \frac{9270}{18} = 515$	

FIGURA 13.6 PANTALLA DE RESULTADOS DE MINITAB PARA EL DISEÑO DE DOS FACTORES PARA EL EXAMEN DE ADMISIÓN

```
SOURCE          DF        SS        MS         F         P
Factor A         2      6100      3050      1.38     0.299
Factor B         2     45300     22650     10.27     0.005
Interaction      4     11200      2800      1.27     0.350
Error            9     19850      2206
Total           17     82450
```

sí hay una diferencia significativa en las medias de las puntuaciones en el examen de admisión. Por último, como el valor-p, 0.350, correspondiente al efecto de la interacción es mayor que $\alpha = 0.05$, no hay un efecto significativo de interacción. Por tanto, en este estudio no se encuentran razones para pensar que los tres programas de preparación difieran en su capacidad de preparación, para este examen de admisión, de estudiantes de las distintas licenciaturas.

Se encontró que la licenciatura sí era un factor significativo. Al revisar los cálculos de la tabla 13.2, se ve que las medias muestrales son: estudiantes de administración $\bar{x}_{.1} = 540$, estudiantes de ingeniería $\bar{x}_{.2} = 560$ y estudiantes de ciencias $\bar{x}_{.3} = 445$. Es posible realizar pruebas para los distintos tratamientos; sin embargo, después de observar las tres medias muestrales es de anticipar que no hay diferencia entre los estudiantes con las licenciaturas de ingeniería y administración. Pero, los estudiantes de ciencias parecen estar menos preparados para este examen que los estudiantes de las otras dos licenciaturas. Quizás esta observación haga que la universidad busque otras opciones para ayudar a estos estudiantes a prepararse para el examen de admisión.

Ejercicios

Métodos



28. En un experimento factorial con dos niveles para el factor A y tres niveles para el factor B se obtuvieron los datos siguientes.

		Factor B		
		Nivel 1	**Nivel 2**	**Nivel 3**
Factor A	**Nivel 1**	135 165	90 66	75 93
	Nivel 2	125 95	127 105	120 136

Realice una prueba para determinar si hay algunos efectos principales significativos o algún efecto de interacción. Use $\alpha = 0.05$.

29. De los cálculos de un experimento factorial con cuatro niveles para el factor A, tres niveles para el factor B y tres replicaciones se obtuvieron los datos siguientes: STC $= 280$, SCA $= 26$, SCB $= 23$, SCAB $= 175$. Dé la tabla ANOVA y pruebe si hay algunos efectos principales significativos o algún efecto de interacción. Use $\alpha = 0.05$.

Aplicaciones

30. Una empresa de venta por catálogo diseñó un experimento factorial para probar el efecto de tamaño y diseño de los anuncios publicitarios, en el catálogo, sobre el número de solicitudes de catálogos recibidas (los datos están dados en miles). Se pusieron a consideración tres diseños publicitarios y dos tamaños. Los datos obtenidos se presentan a continuación. Emplee el ANO-

VA para un diseño factorial para probar si hay efectos significativos debidos al tipo de diseño, al tamaño o a la interacción. Use $\alpha = 0.05$.

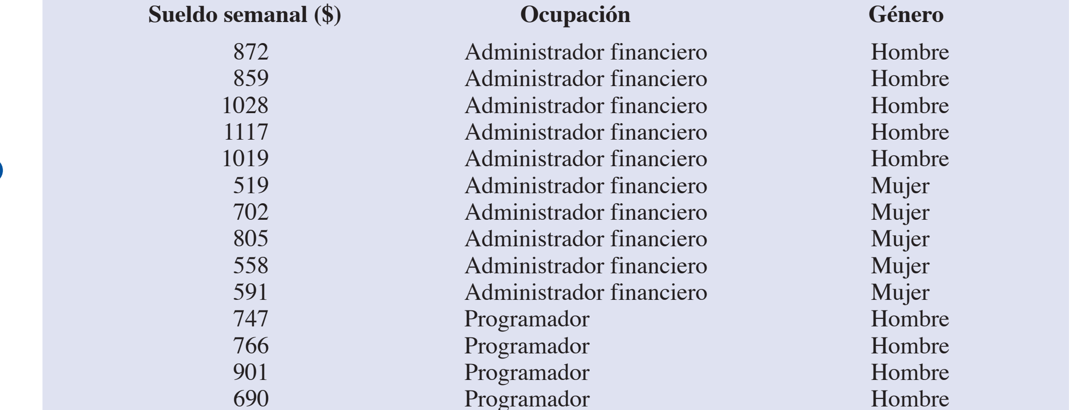
		Tamaño del anuncio	
		Pequeño	**Grande**
Diseño	**A**	8 12	12 8
	B	22 14	26 30
	C	10 18	18 14

31. Un parque de diversión estudió los métodos para disminuir el tiempo de espera (en minutos) al bajar y subir a los pasajeros a los diversos juegos. Se propusieron dos métodos para subir y bajar a los pasajeros de los juegos. Para tomar en cuenta las diferencias debidas al tipo de juego y a la interacción que puede haber entre el tipo de juego y el método de subir y bajar a los pasajeros, se diseñó un experimento factorial. Use los datos siguientes para hacer una prueba sobre cualquier efecto significante debido al método parea subir y bajar a los pasajeros, al tipo de juego y a la interacción. Use $\alpha = 0.05$.

	Tipo de juego		
	Montaña rusa	**Rueda de la fortuna**	**Tobogán**
Método 1	41 43	52 44	50 46
Método 2	49 51	50 46	48 44

32. La U.S. Bureau of Labor Statistics, recoge información sobre sueldos de hombres y mujeres en diversas ocupaciones. Suponga que se desea investigar si hay diferencia entre los sueldos semanales de hombres y mujeres que trabajan como administradores financieros, programadores y farmacéuticos. De cada una de estas ocupaciones se toma una muestra de cinco hombres y cinco mujeres y se registra el sueldo semanal de cada uno de ellos. Los datos obtenidos son los que se presentan a continuación.





Sueldo semanal ($)	Ocupación	Género
872	Administrador financiero	Hombre
859	Administrador financiero	Hombre
1028	Administrador financiero	Hombre
1117	Administrador financiero	Hombre
1019	Administrador financiero	Hombre
519	Administrador financiero	Mujer
702	Administrador financiero	Mujer
805	Administrador financiero	Mujer
558	Administrador financiero	Mujer
591	Administrador financiero	Mujer
747	Programador	Hombre
766	Programador	Hombre
901	Programador	Hombre
690	Programador	Hombre

(*continúa*)

Sueldo semanal ($)	Ocupación	Género
881	Programador	Hombre
884	Programador	Mujer
765	Programador	Mujer
685	Programador	Mujer
700	Programador	Mujer
671	Programador	Mujer
1105	Farmacéutico/a	Hombre
1144	Farmacéutico/a	Hombre
1085	Farmacéutico/a	Hombre
903	Farmacéutico/a	Hombre
998	Farmacéutico/a	Hombre
813	Farmacéutico/a	Mujer
985	Farmacéutico/a	Mujer
1006	Farmacéutico/a	Mujer
1034	Farmacéutico/a	Mujer
817	Farmacéutico/a	Mujer

Utilice el nivel de significancia $\alpha = 0.05$, pruebe si hay algún efecto significativo debido a la ocupación, el género y la interacción.

33. En un estudio publicado en *The Accounting Review* se examinaron los efectos separados y conjuntos de dos grados de presión de tiempo (bajo y moderado) y de tres grados de conocimiento (inexperto, declarativo y de procedimiento) en la conducta al seleccionar palabras clave en una investigación de impuestos. A las personas se les presentaban casos que contenían una serie de hechos, un asunto sobre impuestos y un índice de palabras clave con 1 336 palabras clave. Se les pedía que seleccionaran las palabras clave que creyeran los llevarían a una autoridad tributaria relevante para resolver el caso. Antes del experimento, un grupo de expertos en impuestos determinó que en el texto había 19 palabras clave relevantes. Las personas en el grupo inexperto, poseían poco o ningún conocimiento declarativo o de procedimiento, las personas en el grupo declarativo tenían un conocimiento declarativo significativo pero poco o ningún conocimiento de procedimiento y las personas en el grupo de procedimiento tenían considerables conocimientos declarativos y de procedimiento. El conocimiento declarativo consistía en el conocimiento tanto de las reglas de impuestos aplicables como de los términos técnicos empleados para describir esas reglas. El conocimiento de procedimiento es el conocimiento de las reglas que guían la búsqueda del investigador de impuestos para hallar palabras clave. A las personas en el grupo de poca presión de tiempo se les dieron 25 minutos para resolver un problema, cantidad de tiempo que debía ser "más que adecuada" para resolver el caso, a las personas en el grupo de presión de tiempo moderada se les dieron "sólo" 11 minutos para resolver el caso. Se seleccionaron 25 personas para cada una de las seis combinaciones de tratamientos; las medias muestrales de cada combinación de tratamientos son las que se indican a continuación (las desviaciones estándar están entre paréntesis).

		Conocimiento		
		Inexperto	**Declarativo**	**De procedimiento**
Presión de tiempo	**Baja**	1.13 (1.12)	1.56 (1.33)	2.00 (1.54)
	Moderada	0.48 (0.80)	1.68 (1.36)	2.86 (1.80)

Emplee el ANOVA para probar si hay diferencias significativas debidas a la presión de tiempo, al conocimiento o a la interacción. Use 0.05 como nivel de significancia. La suma total de cuadrados en este experimento fue 327.50.

Resumen

En este capítulo se mostró cómo emplear el análisis de varianza para hallar diferencias entre las medias de varias poblaciones o tratamientos. Se presentó el diseño completamente aleatorizado, el diseño de bloques aleatorizado y el experimento factorial de dos factores. El diseño completamente aleatorizado y el diseño de bloques aleatorizado se usaron para sacar conclusiones acerca de las diferencias en las medias de un solo factor. El objetivo principal de la formación de bloques en el diseño de bloques aleatorizado es eliminar, del término del error, fuentes extrañas de variación. La formación de bloques proporciona una mejor estimación de la verdadera varianza del error y una mejor prueba para determinar si las medias de las poblaciones o tratamientos del factor difieren significativamente.

Se mostró que la base para las pruebas estadísticas empleadas en el análisis de varianza y en el diseño de experimentos es la obtención de dos estimaciones independientes de la varianza poblacional σ^2. En el caso de un solo factor, uno de los estimadores se basa en la variación entre los tratamientos; este estimador sólo proporciona un estimador insesgado de σ^2 si las medias μ_1, μ_2, …, μ_k, son iguales. El otro estimador de σ^2 se basa en la variación de las observaciones dentro de cada muestra; este estimador siempre proporciona un estimador insesgado de σ^2. Al calcular el cociente entre estos dos estimadores (el estadístico F) se obtiene la regla de rechazo para rechazar o no la hipótesis nula que establece que las medias poblacionales o de los tratamientos son iguales. En todos los diseños de experimentos aquí considerados, la partición de las sumas de cuadrados y de los grados de libertad en sus diferentes fuentes permiten calcular las cantidades necesarias para el análisis de varianza y para las pruebas. Se mostró también cómo usar el procedimiento de las LSD de Fisher y el ajuste de Bonferroni para realizar comparaciones por pares y determinar cuáles son las medias que son diferentes.

Glosario

Tabla ANOVA Tabla usada para resumir los cálculos y los resultados del análisis de varianza. Esta tabla tiene columnas en las que se muestran las fuentes de variación, las sumas de cuadrados, los grados de libertad, los cuadrados medios y el o los valores F.

Partición Proceso que distribuye la suma total de cuadrados y de grados de libertad entre sus diversos componentes.

Procedimientos de comparación múltiple Procedimientos estadísticos que se emplean para realizar comparaciones estadísticas entre pares de medias poblacionales.

Tasa de error tipo I por comparación Probabilidad de cometer un error tipo I en la comparación de un solo par.

Tasa de error tipo I por experimentación Probabilidad de cometer un error tipo I en por lo menos una de varias comparaciones por pares.

Factor Otro término empleado para la variable independiente de interés.

Tratamientos Los diferentes niveles (valores) del factor.

Experimento de un solo factor Experimento en el que hay un solo factor con k poblaciones o tratamientos.

Variable de respuesta Otro término para la variable dependiente de interés.

Unidades experimentales Los objetos de interés en el experimento.

Diseño completamente aleatorizado Diseño experimental en el que los tratamientos se asignan en forma aleatoria a las unidades experimentales.

Formación de bloques Proceso que consiste en usar una misma o similares unidades experimentales para todos los tratamientos. El objetivo de la formación de bloques es eliminar, del término del error, fuentes extrañas de variación y con esto proporcionar una prueba más sólida para diferenciar las medias de las poblaciones o tratamientos.

Diseño de bloques aleatorizado Diseño de experimentos en el que se usa la formación de bloques.

Experimento factorial Diseño experimental en el que se obtienen simultáneamente conclusiones acerca de dos o más factores.

Replicaciones Número de veces que en un experimento se repite una condición experimental.
Interacción Efecto que se produce cuando los niveles (valores) de un factor interactúan con los niveles (valores) del otro factor e influyen en la variable de respuesta.

Fórmulas clave

Diseño completamente aleatorizado

Media muestral del tratamiento *j*

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \tag{13.1}$$

Varianza muestral del tratamiento *j*

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1} \tag{13.2}$$

Media muestral general

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n_T} \tag{13.3}$$

$$n_T = n_1 + n_2 + \cdots + n_k \tag{13.4}$$

Cuadrado medio debido a los tratamientos

$$\text{CMTR} = \frac{\text{SCTR}}{k - 1} \tag{13.7}$$

Suma de cuadrados debida a los tratamientos

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 \tag{13.8}$$

Cuadrado medio debido al error

$$\text{CME} = \frac{\text{SCE}}{n_T - k} \tag{13.10}$$

Suma de cuadrados debida al error

$$\text{SCE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 \tag{13.11}$$

Estadístico de prueba para la igualdad de *k* medias poblacionales

$$F = \frac{\text{SCTR}}{\text{CME}} \tag{13.12}$$

Suma de cuadrados del total

$$\text{STC} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\bar{x}})^2 \tag{13.13}$$

Partición de la suma de cuadrados

$$\text{STC} = \text{SCTR} + \text{SCE} \tag{13.14}$$

Procedimiento de comparación múltiple

Prueba estadística para el procedimiento de la LSD de Fisher

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\text{CME}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{13.16}$$

LSD de Fisher

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{CME}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{13.17}$$

Diseño de bloques aleatorizado

Suma de cuadrados del total

$$\text{STC} = \sum_{i=1}^{b}\sum_{j=1}^{k}(x_{ij} - \bar{\bar{x}})^2 \tag{13.22}$$

Suma de cuadrados debida a los tratamientos

$$\text{SCTR} = b\sum_{j=1}^{k}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13.23}$$

Suma de cuadrados debida a los bloques

$$\text{SCBL} = k\sum_{i=1}^{b}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13.24}$$

Suma de cuadrados debida al error

$$\text{SCE} = \text{STC} - \text{SCTR} - \text{SCBL} \tag{13.25}$$

Experimentos factoriales

Suma de cuadrados del total

$$\text{STC} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk} - \bar{\bar{x}})^2 \tag{13.27}$$

Suma de cuadrados del factor A

$$\text{SCA} = br\sum_{i=1}^{a}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13.28}$$

Suma de cuadrados del factor B

$$\text{SCB} = ar\sum_{j=1}^{b}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13.29}$$

Suma de cuadrados debida a la interacción

$$\text{SCAB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2 \tag{13.30}$$

Suma de cuadrados debida al error

$$\text{SCE} = \text{STC} - \text{SCA} - \text{SCB} - \text{SCAB} \tag{13.31}$$

Ejercicios complementarios

34. En un diseño experimental completamente aleatorizado se probó la capacidad absorbente de agua de tres marcas de toallas de papel. Se usaron toallas de un mismo tamaño para probar cuatro secciones de toalla por marca. A continuación se dan los datos de la capacidad de absorción. Emplee 0.05 como nivel de significancia. ¿Parece haber alguna diferencia en la capacidad de absorción de estas marcas?

Marca		
x	*y*	*z*
91	99	83
100	96	88
88	94	89
89	99	76

35. En un estudio publicado en el *Journal of Small Business Management* se concluyó que los individuos que se autoemplean no experimentan tanta satisfacción en el trabajo como los que no se autoemplean. En este estudio, la satisfacción en el trabajo se midió empleando 18 puntos, cada uno de los cuales se evaluaba con una escala de Liker con 1–5 opciones de respuesta que iban de totalmente de acuerdo a totalmente en desacuerdo. En esta escala una puntuación mayor corresponde a mayor satisfacción con el trabajo. La suma de las puntuaciones de los 18 puntos, que iba de 18–90, se empleó para medir la satisfacción con el trabajo. Suponga que se emplea este método para medir la satisfacción en el trabajo de abogados, terapeutas físicos, carpinteros y analistas de sistemas. A continuación se encuentran los resultados obtenidos en una muestra de 10 individuos de cada profesión.



Abogados	Terapeutas físicos	Carpinteros	Analistas de sistemas
44	55	54	44
42	78	65	73
74	80	79	71
42	86	69	60
53	60	79	64
50	59	64	66
45	62	59	41
48	52	78	55
64	55	84	76
38	50	60	62

Con $\alpha = 0.05$ como nivel de significancia, pruebe si hay diferencia en la satisfacción con el trabajo en estas cuatro profesiones.

36. La revista *Money* publicó porcentajes de las proporciones de rendimientos y gastos de acciones y fondos de bonos. Los datos siguientes son las proporciones de gastos en 10 fondos de acciones “midcap”, 10 fondos de acciones “small-cap”, 10 fondos de acciones híbridos y 10 fondos por sector especialista (*Money*, marzo de 2003).

archivo en CD
Funds

Midcap	Small-Cap	Híbridos	Especialista
1.2	2.0	2.0	1.6
1.1	1.2	2.7	2.7
1.0	1.7	1.8	2.6
1.2	1.8	1.5	2.5
1.3	1.5	2.5	1.9
1.8	2.3	1.0	1.5
1.4	1.9	0.9	1.6
1.4	1.3	1.9	2.7
1.0	1.2	1.4	2.2
1.4	1.3	0.3	0.7

Use $\alpha = 0.05$ para probar si hay diferencias significativas entre las proporciones de gastos medios de estos cuatro fondos de acciones.

37. La primera encuesta anual sobre empleo de *Business 2.0* proporcionó datos sobre los salarios anuales de 97 empleos diferentes. Los datos siguientes son salarios anuales de 30 empleos diferentes en tres campos, software y hardware para computadoras, construcción e ingeniería.



Computación		Construcción		Ingeniería	
Empleo	**Salario**	**Empleo**	**Salario**	**Empleo**	**Salario**
Administrador de datos	94	Administrador	55	Aeronáutica	75
Administrador de fabricación	90	Arquitecto	53	Agrícola	70
Programador	63	Administrador arquitecto	77	Química	88
Administrador de proyecto	84	Administrador de la construcción	60	Civil	77
Desarrollador de software	73	Maestro de obras	41	Eléctrica	89
Diseñador	75	Diseñador de interiores	54	Mecánica	85
Sistemas de personal	94	Arquitecto de paisaje	51	Minera	96
Analista de sistemas	77	Estimador	64	Nuclear	105

Use $\alpha = 0.05$ para probar si hay alguna diferencia significativa entre los salarios medios anuales de los tres campos de trabajo.

38. Se proponen tres nuevos métodos de fabricación para un producto nuevo. Para determinar con cuál de los métodos se producen más unidades por hora se elige un diseño experimental completamente aleatorizado y a 30 trabajadores tomados al azar se les asigna alguno de los métodos de fabricación. En la tabla siguiente se presenta el número de unidades producidas por cada uno de los trabajadores.



Método		
A	**B**	**C**
97	93	99
73	100	94
93	93	87
100	55	66
73	77	59
91	91	75
100	85	84
86	73	72
92	90	88
95	83	86

Utilice estos datos y realice una prueba para ver si el número medio de unidades producidas es la misma con los tres métodos de fabricación. Use $\alpha = 0.05$.

39. En un estudio realizado para investigar la actividad de los clientes en las tiendas grandes, a cada cliente se le clasificó al inicio como poco activo, medianamente activo y muy activo. De cada cliente se obtuvo un valor que medía cuán a gusto se encontraba el cliente en la tienda. Valores más altos indicaban que el cliente se encontraba más a gusto. Los datos obtenidos fueron los siguientes.



Poco activo	Medianamente activo	Muy activo
4	5	5
5	6	7
6	5	5
3	4	7
3	7	4
4	4	6
5	6	5
4	5	7

a. Use $\alpha = 0.05$ para probar si hay diferencia en el grado en que se sienten a gusto los tres tipos de clientes.

b. Use el procedimiento de la LSD de Fisher para comparar los grados en que se encuentran a gusto los poco activos y los medianamente activos. Use $\alpha = 0.05$ ¿Cuál es la conclusión?

40. Una empresa realiza una investigación para determinar el rendimiento, en millas por galón, característico de tres marcas de gasolina. Como cada gasolina da rendimientos distintos en automóviles de marcas diferentes, se eligen cinco marcas de automóviles que se tratan como bloques en el experimento; es decir, el automóvil de cada marca se prueba con los tres tipos de gasolina. Los resultados del experimento (en millas por galón) se presentan a continuación.

		Marcas de gasolina		
		I	II	III
Automóviles	A	18	21	20
	B	24	26	27
	C	30	29	34
	D	22	25	24
	E	20	23	24

a. Con $\alpha = 0.05$, ¿se encuentra alguna diferencia entre los rendimientos medios en millas por galón de los tres tipos de gasolina?

b. Analice los datos experimentales usando el ANOVA para diseños completamente aleatorizados. Compare sus hallazgos con los obtenidos en el inciso a. ¿Cuál es la ventaja de tratar de eliminar el efecto de bloque?

41. Wegman's Food Markets y Tops Frendly Markets son cadenas grandes de tiendas de abarrotes en una zona de Nueva York. Cuando Wal-Mart abrió un supermercado en esta zona, los expertos predijeron que Wal-Mart vendería más barato que estas dos tiendas locales. Un periódico publicó los precios de 15 artículos que se presentan en la tabla siguiente.



Artículo	Tops	Wal-Mart	Wegmans
Plátanos (1 lb)	0.49	0.48	0.49
Sopa Cambell's (10.75 oz)	0.60	0.54	0.77
Pechuga de pollo (3 lb)	10.47	8.61	8.07
Pasta de dientes (6.2 oz)	1.99	2.40	1.97
Huevos (1 docena)	1.59	0.88	0.79
Salsa cátsup (36 oz)	2.59	1.78	2.59
Jell-o (3 onz)	0.67	0.42	0.65
Cacahuatina (18 oz)	2.29	1.78	2.09
Leche (descremada, 1/2 gal)	1.34	1.24	1.34
Oscar Meyer hotdogs (1 lb)	3.29	1.50	3.39
Salsa ragú para pasta (1 lb, 10 oz)	2.09	1.50	1.25
Galletas Ritz (1 lb)	3.29	2.00	3.39
Detergente Tid (líquido, 100 oz)	6.79	5.24	5.99
Jugo de naranja Tropicana (1/2 gal)	2.50	2.50	2.50
Twizzlers (frambuesas, 1 lb)	1.19	1.27	1.69

Con $\alpha = 0.05$ como nivel de significancia, pruebe si hay una diferencia significativa entre las tres tiendas en las medias del precio de estos 15 artículos.

42. El U.S. Department of Housing and Urban Development publica datos que muestran el mercado de rentas mensuales en las áreas metropolitanas. Los datos siguientes son las rentas mensuales aceptables en cinco zonas metropolitanas para departamentos de 1, 2 y 3 recámaras (*The New York Times Almanac*, 2006).

	Boston	Miami	San Diego	San Jose	Washington
1 recámara	1077	775	975	1107	1045
2 recámaras	1266	929	1183	1313	1187
3 recámaras	1513	1204	1725	1889	1537

Emplee 0.05 como nivel de significancia, pruebe si las rentas mensuales aceptables son iguales en estas cinco zonas metropolitanas.

43. Se tienen dos sistemas de software para traducción del inglés a otros idiomas. Para ver si hay diferencia en la rapidez de estos dos sistemas de traducción se diseña un experimento factorial. Como el idioma al que se traduzca es también un factor importante, los dos sistemas se prueban traduciendo a tres idiomas: español, francés y alemán. Los datos siguientes dan el tiempo en horas que se necesitó en cada uno de los sistemas.

	Idioma		
	Español	**Francés**	**Alemán**
Sistema 1	8	10	12
	12	14	16
Sistema 2	6	14	16
	10	16	22

Realice una prueba para determinar si hay alguna diferencia significativa de rapidez entre los dos sistemas de software, entre los idiomas a que se traduce y si hay algún efecto de interacción. Use $\alpha = 0.05$.

44. En una fábrica se diseña un experimento factorial para determinar si hay diferencia entre el número de artículos defectuosos producidos por dos máquinas y si el número de defectos depende también de si a estas máquinas se les suministra la materia prima que necesitan, manualmente o mediante un sistema de alimentación automático. A continuación se presentan los datos

del número de artículos defectuosos producidos. Use $\alpha = 0.05$ para probar si hay algún efecto significativo debido a la máquina, al sistema de suministro de la materia prima y a la interacción.

	Suministro de la materia prima	
	Manual	**Automático**
Máquina 1	30	30
	34	26
Máquina 2	20	24
	22	28

Caso problema 1 Centro Médico Wentworth

Como parte de un estudio a largo plazo realizado con personas de 65 años o más, médicos y sociólogos del Centro Médico Wentworth en Nueva York, investigaron la relación entre ubicación geográfica y depresión. Se tomó una muestra de 60 personas, todas en buenas condiciones de salud; 20 de Florida, 20 de Nueva York y 20 de Carolina del Norte. A cada una de estas personas se le aplicó una prueba estandarizada para medir la depresión. Los datos obtenidos se encuentran en el disco compacto en el archivo Medical1.

Otra parte del estudio consistió en analizar la relación entre ubicación geográfica y depresión en personas de 65 años o más pero que tenían algún padecimiento crónico como artritis, hipertensión o padecimientos cardiacos. Para este estudio se tomó también una muestra aleatoria de personas en estas condiciones, 20 de Florida, 20 de Nueva York y 20 de Carolina del Norte. Los datos obtenidos en este estudio se presentan a continuación. Estos datos se encuentran en el disco compacto en el archivo Medical2.





Datos de Medical1			Datos de Medical2		
Florida	**Nueva York**	**Carolina del Norte**	**Florida**	**Nueva York**	**Carolina del Norte**
3	8	10	13	14	10
7	11	7	12	9	12
7	9	3	17	15	15
3	7	5	17	12	18
8	8	11	20	16	12
8	7	8	21	24	14
8	8	4	16	18	17
5	4	3	14	14	8
5	13	7	13	15	14
2	10	8	17	17	16
6	6	8	12	20	18
2	8	7	9	11	17
6	12	3	12	23	19
6	8	9	15	19	15
9	6	8	16	17	13
7	8	12	15	14	14
5	5	6	13	9	11
4	7	3	10	14	12
7	7	8	11	13	13
3	8	11	17	11	11

Informe administrativo

1. Use la estadística descriptiva para resumir los datos de estos dos estudios. ¿Cuáles son sus observaciones preliminares acerca de los valores de depresión?
2. Utilice al análisis de varianza para ambos conjuntos de datos. En cada caso dé las hipótesis a probar. ¿Cuáles son sus conclusiones?
3. Si es necesario use inferencias acerca de las medias de cada uno de los tratamientos. ¿Cuáles son las conclusiones?

Caso problema 2 Compensación para profesionales de ventas

Un grupo local de profesionales de ventas de San Francisco realiza una investigación entre sus miembros para ver si hay alguna relación entre los años de experiencia y el salario de los individuos empleados como vendedores internos y externos. En esta encuesta se pide a los encuestados que especifiquen uno de tres niveles de años de experiencia: bajo (1 a 10 años), medio (11 a 20 años) o alto (21 o más años). A continuación se presenta una parte de los datos obtenidos. El conjunto de datos completo, que contiene 120 observaciones, se encuentra en el archivo SalesSalary del disco compacto que viene con el libro.



Observación	Salario $	Posición	Experiencia
1	53 938	Interno	Medio
2	52 694	Interno	Medio
3	70 515	Externo	Bajo
4	52 031	Interno	Medio
5	62 283	Externo	Bajo
6	57 718	Interno	Bajo
7	79 081	Externo	Alto
8	48 621	Interno	Bajo
9	72 835	Externo	Alto
10	54 768	Interno	Medio
.	.	.	.
.	.	.	.
.	.	.	.
115	58 080	Interno	Alto
116	78 702	Externo	Medio
117	83 131	Externo	Medio
118	57 788	Interno	Alto
119	53 070	Interno	Medio
120	60 259	Externo	Bajo

Informe administrativo

1. Use la estadística descriptiva para resumir los datos.
2. Dé, mediante un intervalo de 95% de confianza, una estimación del salario medio anual de todos los vendedores, sin importar los años de experiencia y el tipo de vendedor.
3. Proporcione, mediante un intervalo de 95% de confianza, una estimación del salario medio anual de los vendedores internos.
4. Dé, mediante un intervalo de 95% de confianza, una estimación del salario medio anual de los vendedores externos.

5. Utilice el análisis de varianza para determinar si hay diferencias significativas debidas al tipo de vendedor (externo o interno). Use 0.05 como nivel de significancia, y, por ahora, ignore el efecto de los años de experiencia.
6. Use el análisis de varianza para determinar si hay diferencias significativas debidas a los años de experiencia. Use 0.05 como nivel de significancia y, por ahora, ignore el efecto del tipo de vendedor (externo o interno).
7. Con 0.05 como nivel de significancia, realice una prueba para determinar si hay diferencias significativas debidas al tipo de vendedor, a los años de experiencia o a la interacción.

Apéndice 13.1 Análisis de varianza con Minitab

Diseño completamente aleatorizado

En la sección 13.2 se mostró el uso del análisis de varianza para probar la igualdad de k medias poblacionales con los datos de un diseño completamente aleatorizado. Para ilustrar el uso de Minitab en este tipo de diseño de experimentos, se muestra cómo probar si las medias del número de unidades producidas semanalmente con cada uno de los métodos del experimento de Chemitech, presentado en la sección 13.1, son iguales. Los datos muestrales se han ingresado en las tres primeras columnas de la hoja de cálculo de Minitab. La columna 1 se rotuló A, la columna 2 se rotuló B y la columna 3 se rotuló C. Mediante los pasos siguientes se obtiene la pantalla de Minitab presentada en la figura 13.5.



Chemitech



Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **ANOVA**
Paso 3. Elegir **One way (Unstacked)**
Paso 4. Cuando aparezca el cuadro de diálogo One-way Analysis of Variance:
Ingresar C1-C3 en el cuadro **Responses** (**in separate columns**)
Clic en **OK**

Diseño de bloques aleatorizado

En la sección 13.4 se mostró el uso del análisis de varianza para probar la igualdad de k medias poblacionales con los datos de un diseño de bloques aleatorizado. Para ilustrar el uso de Minitab en este tipo de diseño de experimentos, se muestra cómo probar si las medias de los grados de estrés de los controladores aéreos es la misma en los tres puestos de trabajo. Los valores de los grados de estrés de la tabla 13.5 se han ingresado en la columna 1 de la hoja de cálculo de Minitab. Codificando los tratamientos como 1, el del sistema A; 2, el del sistema B, y 3, el del sistema C, los valores codificados de los tratamientos se ingresaron en la columna 2 de la hoja de cálculo. Por último el número correspondiente a cada controlador (1, 2, 3, 4, 5 y 6) se ingresó en la columna 3. De esta manera, los valores en el primer renglón de la hoja de cálculo son 15, 1, 1; los valores en el segundo renglón son 15, 2, 1; los valores en el renglón 3 son 18, 3, 1; los valores en el renglón 4 son 14, 1, 2; etc. Con los pasos siguientes se obtiene la pantalla de Minitab que corresponde a la tabla de ANOVA de la tabla 13.8.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **ANOVA**
Paso 3. Elegir **Two-way**
Paso 4. Cuando aparezca el cuadro de diálogo Two-way Analysis of Variance:
Ingresar C1 en el cuadro **Response**
Ingresar C2 en el cuadro **Row factor**
Ingresar C3 en el cuadro **Column factor**
Seleccionar **Fit additive model**
Clic **OK**

Experimento factorial

En la sección 13.5 se mostró el uso del análisis de varianza para probar la igualdad de *k* medias poblacionales con los datos de un experimento factorial. Para ilustrar el uso de Minitab en este tipo de diseño de experimentos, se muestra cómo analizar los datos del experimento de dos factores para el examen de admisión, presentado en esa sección. Las puntuaciones obtenidas en el examen de admisión, que se presentan en la tabla 13.10 se han ingresado en la columna 1 de la hoja de cálculo de Minitab; la columna 1 ha sido rotulada como Puntuación (Score), la columna 2 ha sido rotulada como Factor A, y la columna 3 como Factor B. Para el factor A, los programas de preparación, se ha codificado como 1 el repaso de tres horas, como 2 el programa de un día y como 3 el curso de 10 semanas. Los valores codificados del factor A se han ingresado en la columna 2 de la hoja de cálculo. Para el factor B, las licenciaturas, se ha codificado como 1 administración, como 2 ingeniería y como 3 ciencias. Los valores codificados del factor B se han ingresado en la columna 3. Por tanto, los valores en el primer renglón de la hoja de cálculo son 500, 1, 1; los valores en el renglón 2 son 580, 1, 1; los valores en el renglón 3 son 540, 1, 2; los valores en el renglón 4 son 460, 1, 2; etc. Con los pasos siguientes se obtienen los resultados de Minitab correspondientes a la tabla ANOVA que se muestra en la figura 13.6

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **ANOVA**
Paso 3. Elegir **Two-way**
Paso 4. Cuando aparezca la ventana de diálogo Two-way Analysis of Variance:
Ingresar C1 en el cuadro **Response**
Ingresar C2 en el cuadro **Row factor**
Ingresar C3 en el cuadro **Column factor**
Clic en **OK**

Apéndice 13.2 Análisis de varianza con Excel

Diseño completamente aleatorizado



En la sección 13.2 se mostró el uso del análisis de varianza para probar la igualdad de *k* medias poblacionales con los datos de un diseño completamente aleatorizado. Para ilustrar el uso de Excel en este tipo de diseño de experimentos, se muestra cómo probar si las medias del número de unidades producidas semanalmente con cada uno de los métodos del experimento de Chemitech, presentado en la sección 13.1, son iguales. Los datos muestrales se han ingresado en los renglones 2 a 6 de las columnas A, B y C de la hoja de cálculo de Excel, como se observa en la figura 13.7. Mediante los pasos siguientes se obtienen los resultados que aparecen en las celdas A9:G23; la parte del ANOVA corresponde a la tabla ANOVA presentada en la tabla 13.3.

Paso 1. Seleccionar el menú **Herramientas**
Paso 2. Elegir **Análisis de datos**
Paso 3. Elegir, de la lista Funciones para análisis, **Análisis de varianza de un factor**
Clic en **OK**
Paso 4. Cuando aparezca el cuadro Análisis de varianza de un factor:
Ingresar A1:C6 en el cuadro **Rango de entrada**
Seleccionar **Columnas**
Seleccionar **Rótulos en la primera fila**
Seleccionar **Rango de salida** e ingresar A9 en el cuadro correspondiente
Click en **OK**

FIGURA 13.7 SOLUCIÓN DE EXCEL PARA EL EXPERIMENTO DE CHEMITECH

	A	B	C	D	E	F	G	H
1	**Method A**	**Method B**	**Method C**					
2	58	58	48					
3	64	69	57					
4	55	71	59					
5	66	64	47					
6	67	68	49					
7								
8	Anova: Single Factor							
9								
10	SUMMARY							
11	*Groups*	*Count*	*Sum*	*Average*	*Variance*			
12	Method A	5	310	62	27.5			
13	Method B	5	330	66	26.5			
14	Method C	5	260	52	31			
15								
16								
17	ANOVA							
18	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
19	Between Groups	520	2	260	9.1765	0.0038	3.8853	
20	Within Groups	340	12	28.3333				
21								
22	Total	860	14					
23								
24								

Diseño de bloque aleatorizado



En la sección 13.4 se mostró el uso del análisis de varianza para probar la igualdad de k medias poblacionales con los datos de un diseño de bloques aleatorizado. Para ilustrar el uso de Excel en este tipo de diseño de experimentos, se muestra cómo probar si las medias de los grados de estrés de los controladores aéreos es la misma en los tres puestos de trabajo. Los valores de los grados de estrés de la tabla 13.5 se han ingresado en los renglones 2 a 7 de las columnas B, C y D de la hoja de cálculo, como se observa en la figura 13.8. En las celdas de los renglones 1 a 6 de la columna A se encuentra el número de cada controlador (1, 2, 3, 4, 5, 6). Con los pasos siguientes se obtienen los resultados de Excel y que corresponden a la tabla ANOVA de la tabla 13.8.



Paso 1. Seleccionar el menú **Herramientas**
Paso 2. Elegir **Análisis de datos**
Paso 3. Elegir, de la lista Funciones para análisis, **Análisis de varianza de dos factores con varias muestras por grupo**
Click en **OK**
Paso 4. Cuando aparezca el cuadro de diálogo Análisis de varianza de dos factores sin varias muestras por grupo
Ingresar A1:D7 en el cuadro **Rango de entrada**
Seleccionar **Rótulos**
Seleccionar **Rango de salida** e ingresar A9 en el cuadro
Clic en **OK**

Experimento factorial

En la sección 13.5 se mostró el uso del análisis de varianza para probar la igualdad de k medias poblacionales con los datos de un experimento factorial. Para ilustrar el uso de Excel en este tipo

FIGURA 13.8 SOLUCIÓN DE EXCEL PARA LA PRUEBA DE ESTRÉS DE LOS CONTROLADORES AÉREOS

	A	B	C	D	E	F	G	H
1	**Controller**	**System A**	**System B**	**System C**				
2	1	15	15	18				
3	2	14	14	14				
4	3	10	11	15				
5	4	13	12	17				
6	5	16	13	16				
7	6	13	13	13				
8								
9	Anova: Two-Factor Without Replication							
10								
11	*SUMMARY*	*Count*	*Sum*	*Average*	*Variance*			
12	1	3	48	16	3			
13	2	3	42	14	0			
14	3	3	36	12	7			
15	4	3	42	14	7			
16	5	3	45	15	3			
17	6	3	39	13	0			
18								
19	System A	6	81	13.5	4.3			
20	System B	6	78	13	2			
21	System C	6	93	15.5	3.5			
22								
23								
24	ANOVA							
25	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
26	Rows	30	5	6	3.16	0.0574	3.33	
27	Columns	21	2	10.5	5.53	0.0242	4.10	
28	Error	19	10	1.9				
29								
30	Total	70	17					
31								



de diseño de experimentos, se muestra cómo analizar los datos del experimento de dos factores para el examen de admisión, presentado en esa sección. Las puntuaciones obtenidas en el examen de admisión, que se presentan en la tabla 13.10 se han ingresado en los renglones 2 a 7 de las columnas B, C y D de la hoja de cálculo como se observa en la figura 13.9. Con los pasos siguientes se muestran los resultados que se observan en las celdas A9:G44; la parte del ANOVA corresponde a la tabla de ANOVA de la figura 13.6.

Paso 1. Seleccionar el menú **Herramientas**

Paso 2. Elegir **Análisis de datos**

Paso 3. Elegir, de la lista de Funciones para análisis, **Análisis de varianza de dos factores con varias muestras por grupo**

Click en **OK**

Paso 4. Cuando aparezca el cuadro de diálogo Análisis de varianza de dos factores con varias muestras por grupo:

Ingresar A1:D7 en el cuadro **Rango de entrada**

Ingresar 2 en el cuadro **Fila por muestra**

Seleccionar **Rango de salida** e ingresar A9 en el cuadro correspondiente

Click en **OK**

FIGURA 13.9 SOLUCIÓN DE EXCEL PARA EL EXPERIMENTO DE DOS FACTORES DEL EXAMEN DE ADMISIÓN

	A	B	C	D	E	F	G	H
1		**Business**	**Engineering**	**Arts and Sciences**				
2	**3-hour review**	500	540	480				
3		580	460	400				
4	**1-day program**	460	560	420				
5		540	620	480				
6	**10-week course**	560	600	480				
7		600	580	410				
8								
9	Anova: Two-Factor With Replication							
10								
11	SUMMARY	Business	Engineering	Arts and Sciences	Total			
12	*3-hour review*							
13	Count	2	2	2	6			
14	Sum	1080	1000	880	2960			
15	Average	540	500	440	493.33333			
16	Variance	3200	3200	3200	3946.6667			
17								
18	*1-day program*							
19	Count	2	2	2	6			
20	Sum	1000	1180	900	3080			
21	Average	500	590	450	513.33333			
22	Variance	3200	1800	1800	5386.6667			
23								
24	*10-week course*							
25	Count	2	2	2	6			
26	Sum	1160	1180	890	3230			
27	Average	580	590	445	538.33333			
28	Variance	800	200	2450	5936.6667			
29								
30	*Total*							
31	Count	6	6	6				
32	Sum	3240	3360	2670				
33	Average	540	560	445				
34	Variance	2720	3200	1510				
35								
36								
37	ANOVA							
38	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
39	Sample	6100	2	3050	1.38	0.2994	4.26	
40	Columns	45300	2	22650	10.27	0.0048	4.26	
41	Interaction	11200	4	2800	1.27	0.3503	3.63	
42	Within	19850	9	2205.5556				
43								
44	Total	82450	17					
45								

CAPÍTULO 14

Regresión lineal simple

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

ALLIANCE DATA SYSTEMS*
DALLAS TEXAS

Alliance Data Systems (ADS), una empresa de la creciente industria de administración de la relación con el cliente (Customer Relationship Management, CRM) proporciona servicios de transacciones, crédito y mercadotecnia. Los clientes de ADS están concentrados en cuatro industrias: industria minorista, supermercados pequeños, derivados del petróleo/energía eléctrica y transporte. En 1983, Alliance empezó ofreciendo servicios extremo a extremo de tramitación de crédito para la industria minorista, la industria de derivados del petróleo y la industria de restaurantes de categoría media: actualmente ADS emplea a más de 6500 personas que proporcionan servicios a clientes en todo el mundo. ADS, sólo en Estados Unidos, opera más de 140 000 terminales de punto de venta, y procesa más de 2.5 miles de millones de transacciones anuales. En Estados Unidos ADS es la segunda empresa en servicios de crédito de establecimientos locales representando 49 programas de establecimientos locales con casi 72 millones de tarjetahabientes. En 2001, ADS hizo una oferta pública inicial y ahora cotiza en la bolsa de Nueva York.

Uno de los servicios de mercadotecnia ofrecidos por ADS es el de campañas y publicidad directas por correo. La empresa posee una base de datos con información sobre los hábitos de consumo de más de 100 millones de consumidores, lo que le permite dirigir sus acciones a los consumidores que tienen la mayor probabilidad de beneficiarse de la publicidad por correo. El Grupo de desarrollo analítico de ADS emplea el análisis de regresión en la obtención de modelos para medir y predecir la receptividad del consumidor a las campañas de mercadotecnia directa. Algunos modelos de regresión predicen la probabilidad de compra de las personas que reciben la publicidad y otros predicen la cantidad que gastarán cuando realicen una compra.

En una determinada campaña, una cadena de tiendas deseaba atraer a nuevos clientes. Para predecir el efecto de la campaña, los analistas de ADS tomaron de la base de datos una muestra de consumidores, les enviaron material promocional y después recogieron datos sobre la respuesta de los consumidores. Los datos recogidos se referían al monto de la compra realizada por los consumidores que respondieron a la campaña, así como a diversas variables específicas del consumidor, que se consideraron útiles para predecir las ventas. La variable del consumidor que más contribuyó a predecir el monto de compra fue la cantidad total de compras a crédito realizadas en tiendas semejantes en los últimos 39 meses. Los analistas de ADS obtuvieron una ecuación de regresión estimada con la que se relacionaba el monto de compra con la cantidad gastada en tiendas semejantes:

$$\hat{y} = 26.7 + 0.00205x$$

donde

$\hat{y}$ = monto de compra
x = monto gastado en tiendas similares

Con esta ecuación, pudieron predecir que una persona que hubiera gastado $10 000 en tiendas semejantes en los últimos 39 meses, gastaría $47.20 como respuesta a la publicidad directa por correo. En este capítulo se verá cómo obtener estas ecuaciones de regresión estimada.

En el modelo final que obtuvieron los analistas de ADS también participaban algunas otras variables que incrementaban el poder predictivo de la ecuación de predicción. Entre estas variables se encontraba la existencia o no de una tarjeta de crédito, el ingreso estimado, y la cantidad promedio gastada en cada visita a la tienda seleccionada. En el capítulo siguiente se verá cómo incorporar estas variables adicionales a un modelo de regresión múltiple.



Analistas de ADS discuten sobre el uso del modelo de regresión para predecir las ventas en una campaña de comercialización directa. © Cortesía de Alliance Data Systems.

*Los autores agradecemos a Philip Clemance de Desarrollo analítico de Alliance Data Systems por proporcionarnos este artículo para *La estadística en la práctica*.

En la administración, las decisiones suelen basarse en la relación entre dos o más variables. Por ejemplo, observar la relación entre el gasto en publicidad y las ventas puede permitir a un gerente de mercadotecnia tratar de predecir las ventas correspondientes a un determinado gasto en publicidad. O, una empresa de servicios públicos puede emplear la relación entre la temperatura diaria y la demanda de electricidad para predecir la demanda de electricidad considerando las temperaturas diarias que se esperan el mes siguiente. Algunas veces los directivos se apoyan en la intuición para juzgar la relación entre dos variables. Sin embargo, cuando es posible tener datos, puede emplearse un procedimiento estadístico llamado *análisis de regresión* para obtener una ecuación que indique cuál es la relación entre las variables.

Sir Francis Galton (1822-1911) fue el primero en emplear los métodos estadísticos para estudiar la relación entre dos variables. Galton estaba interesado en estudiar la relación entre la estatura de padre e hijo. Karl Pearson (1857-1936) analizó esta relación en 1078 pares de padre-hijo.

En la terminología que se emplea en regresión, a la variable que se va a predecir se le llama **variable dependiente**. A la variable o variables que se usan para predecir el valor de la variable dependiente se les llama **variables independientes**. Por ejemplo, al analizar el efecto de los gastos en publicidad sobre las ventas, como lo que busca el gerente de mercadotecnia es predecir las ventas, esto indica que las ventas serán la variable dependiente.

En este capítulo se estudia el tipo más sencillo de análisis de regresión en el que interviene una variable independiente y una variable dependiente y en el que la relación entre estas variables es aproximada mediante una línea recta. A este tipo de análisis de regresión se le conoce como **regresión lineal simple**. Al análisis de regresión en el que intervienen dos o más variables independientes se le llama análisis de regresión múltiple; el análisis de regresión múltiple y los casos en los que la relación es curvilínea se estudian en los capítulos 15 y 16.

14.1 Modelo de regresión lineal simple

Armand's Pizza Parlors es una cadena de restaurantes de comida italiana. Sus mejores ubicaciones son las que se encuentran cerca de los campus de las universidades. Los gerentes creen que las ventas trimestrales de estos restaurantes (que se denotan por y) están directamente relacionadas con el tamaño de la población estudiantil (que se denota x); es decir, en los restaurantes que están cerca de campus que tienen una población estudiantil grande se generan más ventas que en los restaurantes situados cerca de campus con una población estudiantil pequeña. Empleando el análisis de regresión, se puede obtener una ecuación que muestre cuál es la relación entre la variable dependiente y y la variable dependiente x.

Modelo de regresión y ecuación de regresión

En el ejemplo de los restaurantes Armand's Pizza Parlors, la población consta de todos los restaurantes Armand. Para cada restaurante de la población, hay un valor x (población estudiantil) y un correspondiente valor y (ventas trimestrales). A la ecuación con que se describe cómo se relaciona y con x y en la que se da un término para el error, se le llama **modelo de regresión**. El siguiente es el modelo que se emplea en la regresión lineal simple.

MODELO DE REGRESIÓN LINEAL SIMPLE

$$y = \beta_0 + \beta_1 x + \epsilon \qquad \textbf{(14.1)}$$

β_0 y β_1 se conocen como los parámetros del modelo, y ϵ (la letra griega épsilón) es una variable aleatoria que se conoce como término del error. El término del error da cuenta de la variabilidad de y que no puede ser explicada por la relación lineal entre x y y.

La población de los restaurantes Armand's puede verse también como una colección de subpoblaciones, una para cada uno de los valores de x. Por ejemplo, una subpoblación está formada por todos los campus universitarios de 8000 estudiantes; otra subpoblación consta de todos los restaurantes Armand's localizados cerca de los campus universitarios de 9000 estudiantes; etc. Para cada subpoblación hay una distribución de valores y. Así, hay una distribución de valores y que corresponde a los restaurantes localizados cerca de los campus de 8000 estudiantes; hay otra distribución de valores y que corresponde a los restaurantes ubicados cerca de los campus de 9000 estudiantes, y así sucesivamente. Cada una de estas distribuciones de valores y tiene su propia media o valor esperado. A la ecuación que describe la relación entre el valor esperado de y, que se denota $E(x)$, y x se le llama **ecuación de regresión**. La siguiente es la ecuación de regresión para la regresión lineal simple.

ECUACIÓN DE REGRESIÓN LINEAL SIMPLE

$$E(y) = \beta_0 + \beta_1 x \tag{14.2}$$

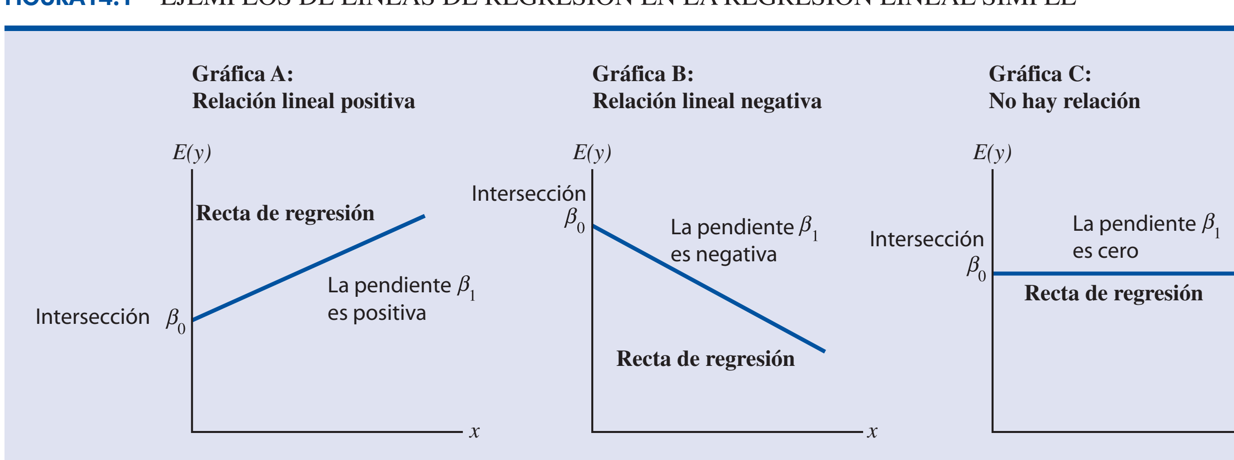
La gráfica de la ecuación de regresión lineal simple es una línea recta; β_0 es la intersección de la recta de regresión con el eje y, β_1 es la pendiente y $E(y)$ es la media o valor esperado de y para un valor dado de x.

En la figura 14.1 se presentan ejemplos de posibles rectas de regresión. La recta de regresión de la gráfica A indica que el valor medio de y está relacionado positivamente con x. La recta de regresión de la gráfica B indica que el valor medio de y está relacionado negativamente con x, valores menores de $E(y)$ corresponden a valores mayores de x. La recta de regresión de la gráfica C muestra el caso en el que el valor medio de y no está relacionado con x; es decir, el valor medio de y es el mismo para todos los valores de x.

Ecuación de regresión estimada

Si se conocieran los valores de los parámetros poblacionales β_0 y β_1, se podría emplear la ecuación (14.2) para calcular el valor medio de y para un valor dado de x. Sin embargo, en la práctica no se conocen los valores de estos parámetros y es necesario estimarlos usando datos muestrales. Se calculan estadísticos muestrales (que se denotan b_0 y b_1) como estimaciones de los parámetros poblacionales β_0 y β_1. Sustituyendo en la ecuación de regresión b_0 y b_1 por los

FIGURA 14.1 EJEMPLOS DE LÍNEAS DE REGRESIÓN EN LA REGRESIÓN LINEAL SIMPLE



valores de los estadísticos muestrales β_0 y β_1, se obtiene la **ecuación de regresión estimada**. La ecuación de regresión estimada de la regresión lineal simple se da a continuación.

ECUACIÓN DE REGRESIÓN LINEAL SIMPLE ESTIMADA

$$\hat{y} = b_0 + b_1x \tag{14.3}$$

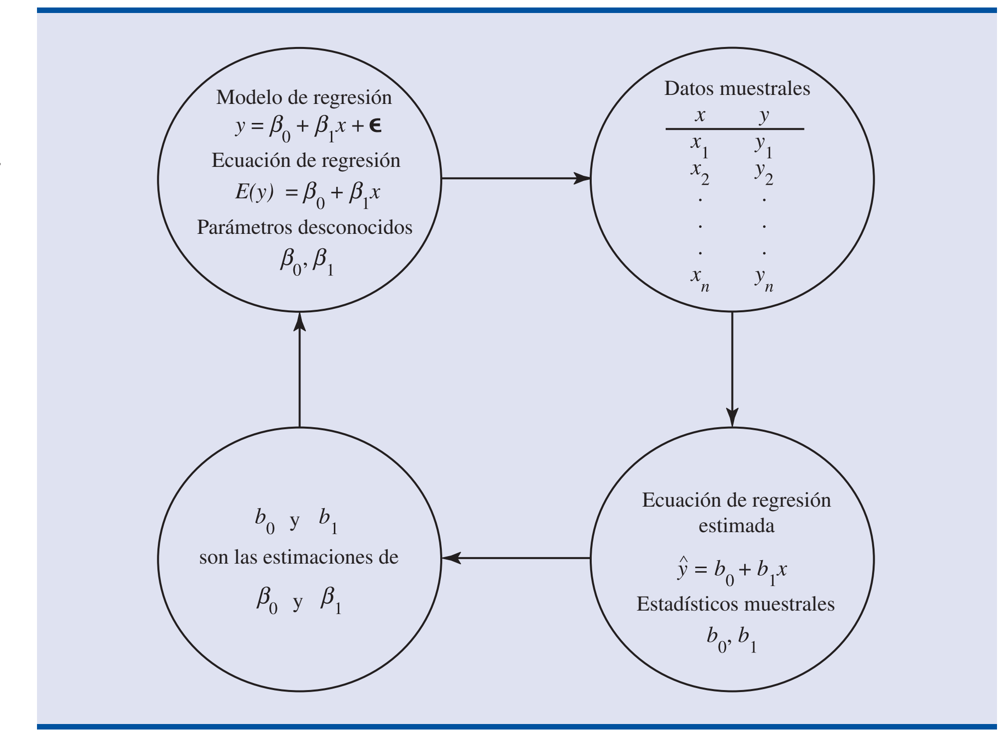
A la gráfica de la ecuación de regresión simple estimada se le llama *recta de regresión estimada*; b_0 es la intersección con el eje y y b_1 es la pendiente. En la sección siguiente se muestra el uso del método de mínimos cuadrados para calcular los valores de b_0 y b_1 para la ecuación de regresión estimada.

En general, $\hat{y}$ es el estimador puntual de $E(y)$, el valor medio de las y para un valor dado de x. Por lo tanto, para estimar la media o el valor esperado de las ventas trimestrales de todos los restaurantes situados cerca de los campus de 10 000 estudiantes, Armad's tendrá que sustituir en la ecuación (14.3) x por 10 000. pero, en algunos casos, a Armand's lo que le interesará será predecir las ventas de un determinado restaurante. Por ejemplo, supóngase que Armand's desee predecir las ventas trimestrales del restaurante que se encuentra cerca de Talbot Collage, una escuela de 10 000 estudiantes. Resulta que la mejor estimación de la y que corresponde a un determinado valor de x es también la proporcionada por $\hat{y}$. Por lo tanto, para predecir las ventas trimestrales del restaurante ubicado cerca de Talbot Collage, Armad's también sustituirá la x de la ecuación (14.3) por 10 000.

Como el valor de $\hat{y}$ proporciona tanto una estimación puntual de $E(x)$ para un valor dado de x como una estimación puntual de un solo valor de y para un valor dado de x, a $\hat{y}$ se le llamará simplemente *valor estimado de y*. En la figura 14.2 se presenta en forma resumida el proceso de estimación en la regresión lineal simple.

La estimación de β_0 y β_1 es un proceso estadístico muy parecido a la estimación de μ que se vio en el capítulo 7. β_0 y β_1 son los parámetros de interés que son desconocidos, y b_0 y b_1 son los estadísticos muestrales que se usan para estimar los parámetros.

FIGURA 14.2 PROCESO DE ESTIMACIÓN EN LA REGRESIÓN LINEAL SIMPLE



NOTAS Y COMENTARIOS

1. El análisis de regresión no puede entenderse como un procedimiento para establecer una relación de causa y efecto entre las variables. Este procedimiento sólo indica cómo o en qué medida las variables están relacionadas una con otra. Conclusiones acerca de una relación causa y efecto deben basarse en los conocimientos de los especialistas en la aplicación de que se trate.

2. La ecuación de regresión en la regresión lineal simple es $E(y) = \beta_0 + \beta_1 x$. En libros más avanzados sobre análisis de regresión se suele escribir la ecuación de regresión como $E(y|x) = \beta_0 + \beta_1 x$ enfatizando así que lo que proporciona esta ecuación es el valor medio de las y para un valor dado de x.



14.2 Método de mínimos cuadrados

En la regresión lineal simple, cada observación consta de dos valores: uno de la variable independiente y otro de la variable dependiente.

El **método de mínimos cuadrados** es un método en el que se usan los datos muestrales para hallar la ecuación de regresión estimada. Para ilustrar el método de mínimos cuadrados, supóngase que se recolectan datos de una muestra de 10 restaurantes Armand's Pizza Parlors ubicados todos cerca de campus universitarios. Para la observación i o el restaurante i de la muestra, x_i es el tamaño de la población de estudiantes (en miles) en el campus y y_i son las ventas trimestrales (en miles de dólares). En la tabla 14.1 se presentan los valores de x_i y y_i en esta muestra de 10 restaurantes. Como se ve, el restaurante 1, para el que $x_1 = 2$ y $y_1 = 58$, está cerca de un campus de 2000 estudiantes y sus ventas trimestrales son de $58 000. El restaurante 2, para el que $x_2 = 6$ y $y_2 = 105$, está cerca de un campus de 6000 estudiantes y sus ventas trimestrales son de $105 000. El valor mayor es el que corresponde a ventas del restaurante 10, el cual está cerca de un campus de 26 000 estudiantes y sus ventas trimestrales son de $202 000.

La figura 14.3 es el diagrama de dispersión de los datos de la tabla 14.1. La población de estudiantes se indica en el eje horizontal y las ventas trimestrales en el eje vertical. Los **diagramas de dispersión** para el análisis de regresión se trazan colocando la variable independiente x en el eje horizontal y la variable dependiente y en el eje vertical. El diagrama de dispersión permite observar gráficamente los datos y obtener conclusiones acerca de la relación entre las variables.

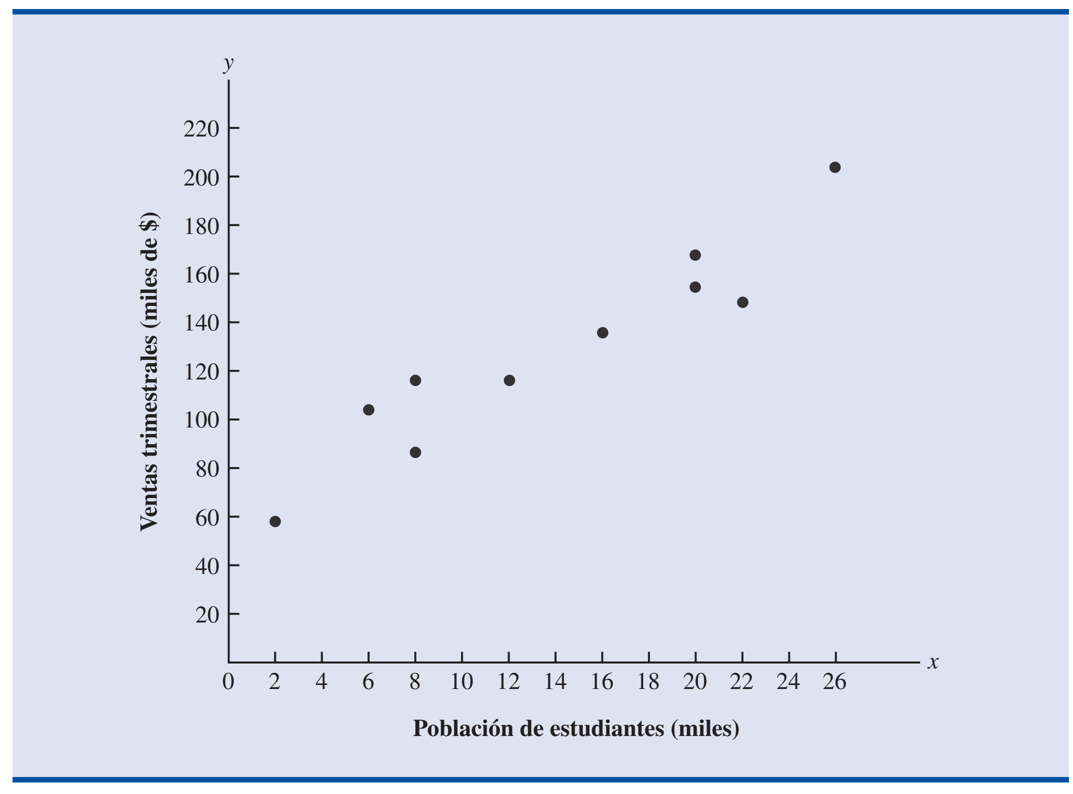
¿Qué conclusión preliminar se puede obtener de la figura 14.3? Las ventas trimestrales parecen ser mayores cerca de campus en los que la población de estudiantes es mayor. Además, en estos datos se observa que la relación entre el tamaño de la población de estudiantes y las ventas trimestrales parece poder aproximarse mediante una línea recta; en efecto, se observa que hay

TABLA 14.1 POBLACIÓN DE ESTUDIANTES Y VENTAS TRIMESTRALES EN 10 RESTAURANTES ARMAND'S PIZZA PARLORS



Restaurante i	Población de estudiantes (miles) x_i	Ventas trimestrales (miles de $) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

FIGURA 14.3 DIAGRAMA DE DISPERSIÓN EN EL QUE SE MUESTRA LA POBLACIÓN DE ESTUDIANTES Y LAS VENTAS TRIMESTRALES DE ARMAND'S PIZZA PARLORS



una relación lineal positiva entre x y y. Por tanto, para representar la relación entre ventas trimestrales y la población de estudiantes, se elige el modelo de regresión lineal simple. Decidido esto, la tarea siguiente es usar los datos muestrales de la tabla 14.1 para determinar los valores de b_0 y b_1 en la ecuación de regresión lineal simple. Para el restaurante i, la ecuación de regresión simple estimada es

$$\hat{y}_i = b_0 + b_1 x_i \tag{14.4}$$

donde

$\hat{y}_i$ = valor estimado de las ventas trimestrales (en miles de dólares) del restaurante i
b_0 = intersección de la recta de regresión con el eje y
b_1 = pendiente de la recta de regresión
x_i = tamaño de la población de estudiantes (en miles) del restaurante i

Como para el restaurante i, y_i denota ventas observadas (reales) y $\hat{y}_i$ denota ventas estimadas mediante la ecuación (14.4), para cada uno de los restaurantes de la muestra habrá un valor de ventas observadas y_i y un valor de ventas estimadas $\hat{y}_i$. Para que la recta de regresión estimada proporcione un buen ajuste a los datos, las diferencias entre los valores observados y los valores estimados deben ser pequeñas.

En el método de mínimos cuadrados se usan los datos muestrales para obtener los valores de b_0 y b_1 que minimicen la *suma de los cuadrados de las desviaciones (diferencias)* entre los valores observados de la variable dependiente y_i y los valores estimados de la variable dependiente. El criterio que se emplea en el método de mínimos cuadrados es el de la expresión (14.5).

Carl Friedrich Gauss (1777- 1855) fue quien propuso el método de mínimos cuadrados.

CRITERIO DE MÍNIMOS CUADRADOS

$$\min \Sigma(y_i - \hat{y}_i)^2 \tag{14.5}$$

donde

y_i = valor observado de la variable dependiente en la observación i
$\hat{y}_i$ = valor estimado de la variable independiente en la observación i

Se puede usar cálculos diferenciales para demostrar (véase apéndice 14.1) que los valores de b_0 y b_1 que minimiza la expresión (14.5) se pueden encontrar usando las ecuaciones (14.6) y (14.7).

Al calcular b_1 con una calculadora, en los cálculos intermedios deben llevarse tantas cifras significativas como sea posible. Se recomienda llevar por lo menos cuatro cifras significativas

PENDIENTE E INTERSECCIÓN CON EL EJE y DE LA ECUACIÓN DE REGRESIÓN ESTIMADA*

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \tag{14.6}$$

$$b_0 = \bar{y} - b_1\bar{x} \tag{14.7}$$

donde

x_i = valor de la variable independiente en la observación i
y_i = valor de la variable dependiente en la observación i
$\bar{x}$ = media de la variable independiente
$\bar{y}$ = media de la variable dependiente
n = número total de observaciones

En la tabla 14.2 se presentan los cálculos necesarios para obtener la ecuación de regresión estimada en el ejemplo de Armand's Pizza Parlors. Como la muestra es de 10 restaurantes, tenemos 10 observaciones. Dado que en las ecuaciones (14.6) y (14.7) se necesitan $\bar{x}$ y $\bar{y}$, se empieza por calcular $\bar{x}$ y $\bar{y}$.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{1300}{10} = 130$$

Usando las ecuaciones (14.6) y (14.7) y la información de la tabla 14.2, se calcula la pendiente y la intersección con el eje y de la ecuación de regresión de Armand's Pizza Parlors. La pendiente (b_1) se calcula como sigue.

*Otra fórmula de calcular b_1, es

$$b_1 = \frac{\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n}$$

Esta forma de la ecuación (14.6) suele recomendarse cuando se emplea una calculadora para calcular b_1.

TABLA 14.2 ECUACIÓN DE REGRESIÓN ESTIMADA PARA ARMAND'S PIZZA PARLORS OBTENIDA POR EL MÉTODO DE MÍNIMOS CUADRADOS

Restaurante i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totales	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

La intersección con el eje y (b_0) se calcula como sigue.

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

Por lo tanto, la ecuación de regresión estimada es

$$\hat{y} = 60 + 5x$$

En la figura 14.4 se muestra esta ecuación graficada sobre el diagrama de dispersión.

La pendiente de la ecuación de regresión estimada ($b_1 = 5$) es positiva, lo que implica que a medida que aumenta el tamaño de la población de estudiantes, aumentan las ventas. Se concluye (basándose en las ventas dadas en miles de $ y en el tamaño de la población de estudiantes en miles) que un aumento de 1000 en el tamaño de la población de estudiantes corresponde a un aumento esperado de $5000 en las ventas; es decir, se espera que las ventas trimestrales aumenten $5 por cada aumento de un estudiante.

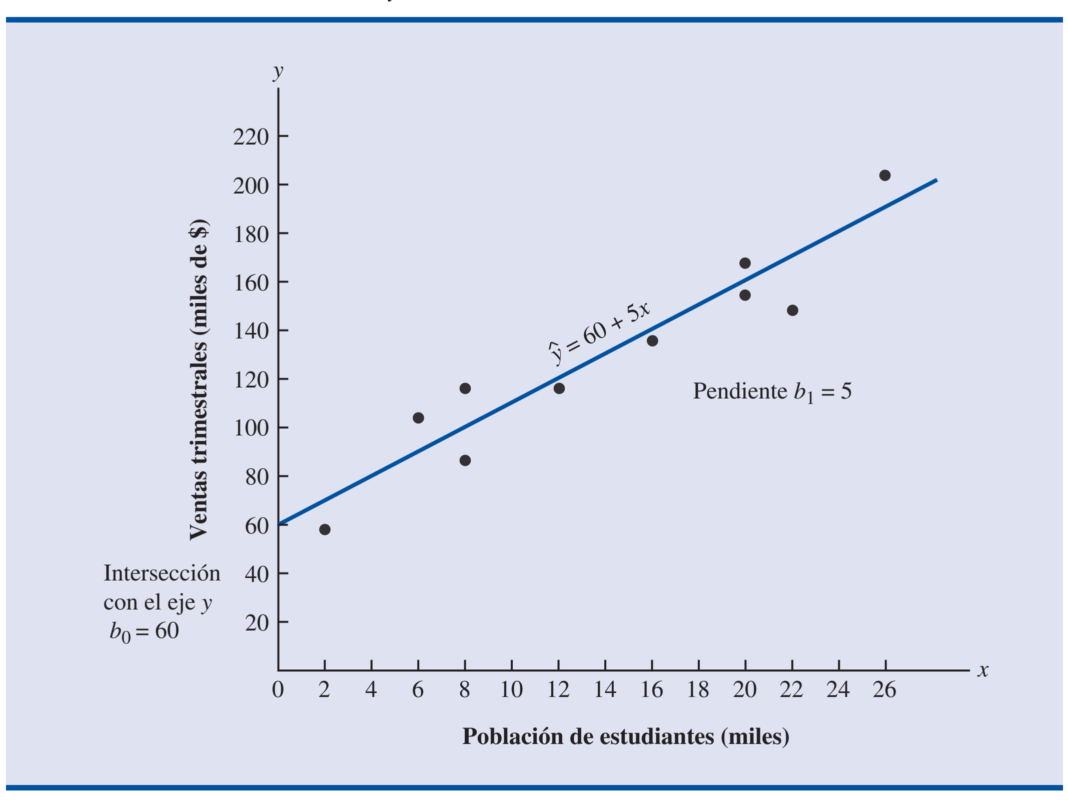
Debe tenerse mucho cuidado al usar la ecuación de regresión estimada para hacer predicciones fuera del rango de valores de la variable independiente, ya que fuera de ese rango no puede asegurarse que esta relación sea válida.

Si se considera que la ecuación de regresión estimada obtenida por el método de mínimos cuadrados describe adecuadamente la relación entre x y y, parecerá razonable usar esta ecuación de regresión estimada para estimar el valor de y para un valor dado de x. Por ejemplo, si se quisieran predecir las ventas trimestrales de un restaurante ubicado cerca de un campus de 16 000 estudiantes, se calcularía

$$\hat{y} = 60 + 5(16) = 140$$

De manera que las ventas trimestrales pronosticadas para este restaurante serían de $140 000. En la sección siguiente se verán los métodos para evaluar el uso correcto de la ecuación de regresión para hacer estimaciones y predicciones.

FIGURA 14.4 GRÁFICA DE LA ECUACIÓN DE REGRESIÓN ESTIMADA DE ARMAND'S PIZZA PARLORS: $\hat{y} = 60 + 5x$



NOTAS Y COMENTARIOS

El método de mínimos cuadrados proporciona una ecuación de regresión estimada que minimiza la suma de los cuadrados de las desviaciones entre los valores observados de la variable dependiente y_i y los valores estimados de la variable dependiente $\hat{y}_i$. El criterio de mínimos cuadrados permite obtener la ecuación de mejor ajuste. Si se empleara otro criterio, como minimizar la suma de las desviaciones absolutas entre y_i y $\hat{y}_i$, se obtendría una ecuación diferente. En la práctica el método de mínimos cuadrados es el método más usado.

Ejercicios

Método



1. Dadas las siguientes cinco observaciones de las variables x y y.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Trace el diagrama de dispersión correspondiente a estos datos.
b. ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre las dos variables?

c. Trate de aproximar la relación entre x y y trazando una línea recta que pase a través de los puntos de los datos.
d. Con las ecuaciones (14.6) y (14.7) calcule b_0 y b_1 para obtener la ecuación de regresión estimada.
e. Use la ecuación de regresión estimada para predecir el valor de y cuando $x = 4$.

2. Dadas las siguientes cinco observaciones de las variables x y y.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a. Trace, con estos datos, el diagrama de dispersión.
b. ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre las dos variables?
c. Trate de aproximar la relación entre x y y trazando una línea recta a través de los puntos de los datos.
d. Con las ecuaciones (14.6) y (14.7) calcule b_0 y b_1, para obtener la ecuación de regresión estimada.
e. Use la ecuación de regresión estimada para predecir el valor de y cuando $x = 4$.

3. Dadas las observaciones siguientes sobre estas dos variables obtenidas en un estudio de regresión.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

a. Con estos datos trace el diagrama de dispersión.
b. Obtenga la ecuación de regresión estimada correspondiente a estos datos.
c. Use la ecuación de regresión estimada para predecir el valor de y cuando $x = 4$.

Aplicaciones





4. Los datos siguientes son estaturas y pesos de nadadoras.



Estatura	68	64	62	65	66
Peso	132	108	102	115	128

a. Trace el diagrama de dispersión de estos datos usando la estatura como variable independiente.
b. ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre las dos variables?
c. Trate de aproximar la relación entre estatura y peso trazando una línea recta a través de los puntos de los datos.
d. Obtenga la ecuación de regresión estimada calculando b_0 y b_1
e. Si la estatura de una nadadora es 63 pulgadas, ¿cuál será su peso estimado?

5. Los adelantos tecnológicos han hecho posible fabricar botes inflables. Estos botes de goma inflables, que pueden enrollarse formando un paquete no mayor que una bolsa de golf, tienen tamaño suficiente para dos pasajeros con su equipo de excursionismo. La revista *Canoe & Kayac* probó los botes de nueve fabricantes para ver su funcionamiento en un recorrido de tres días. Uno de los criterios de evaluación fue su capacidad para equipaje que se evaluó utilizando una escala de 4 puntos, siendo 1 la puntuación más baja y 4 la puntuación más alta. Los datos siguientes muestran la evaluación que obtuvieron respecto a capacidad para equipaje y los precios de los botes (*Canoe Kayak,* marzo 2003).



Bote	Capacidad para equipaje	Precio ($)
S14	4	1595
Orinoco	4	1399
Outside Pro	4	1890
Explorer 380X	3	795
River XK2	2.5	600
Sea Tiger	4	1995
Maverik II	3	1205
Starlite 100	2	583
Fat Pack Cat	3	1048

a. Trace el diagrama de dispersión de estos datos empleando la capacidad para equipaje como variable independiente.
b. ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre capacidad para equipaje y precio?
c. A través de los puntos de los datos trace una línea recta para aproximar la relación lineal entre capacidad para equipaje y precio.
d. Utilice el método de mínimos cuadrados para obtener la ecuación de regresión estimada.
e. Dé una interpretación de la pendiente de la ecuación de regresión estimada.
f. Diga cuál será el precio de un bote que tenga 3 en la evaluación de su capacidad para equipaje.

6. Wageweb realiza estudios sobre datos salariales y presenta resúmenes de éstos en su sitio de la Red. Basándose en datos salariales desde el 1 de octubre de 2002 Wageweb publicó que el salario anual promedio de los vicepresidentes de ventas era $142 111 con una gratificación anual promedio de $15 432 (Wageweb.com, 13 de mazo de 2003). Suponga que los datos siguientes sean una muestra de salarios y bonos anuales de 10 vicepresidentes de ventas. Los datos se dan en miles de dólares.



Vicepresidente	Salario	Gratificación
1	135	12
2	115	14
3	146	16
4	167	19
5	165	22
6	176	24
7	98	7
8	136	17
9	163	18
10	119	11

a. Trace un diagrama de dispersión con estos datos tomando como variable independiente los salarios.
b. ¿Qué indica el diagrama de dispersión del inciso a) acerca de la relación entre salario y gratificación?
c. Use el método de mínimos cuadrados para obtener la ecuación de regresión estimada.
d. Dé una interpretación de la ecuación de regresión estimada.
e. ¿Cuál será la gratificación de un vicepresidente que tenga un salario anual de $120 000?

7. ¿Esperaría que los automóviles más confiables fueran los más caros? *Consumer Reports* evaluó 15 de los mejores automóviles sedán. La confiabilidad se evaluó con una escala de 5 puntos: mala (1), regular (2), buena (3), muy buena (4) y excelente (5). Los precios y la evaluación sobre la confiabilidad de estos 15 automóviles se presenta en la tabla siguiente (*Consumer Reports,* febrero de 2004).



Marca y modelo	Confiabilidad	Precio ($)
Acura TL	4	33 150
BMW 330i	3	40 570
Lexus IS300	5	35 105
Lexus ES330	5	35 174
Mercedes-Benz C320	1	42 230
Lincoln LS Premium (V6)	3	38 225
Audi A4 3.0 Quattro	2	37 605
Cadillac CTS	1	37 695
Nissan Maxima 3.5 SE	4	34 390
Infiniti I35	5	33 845
Saab 9-3 Aero	3	36 910
Infiniti G35	4	34 695
Jaguar X-Type 3.0	1	37 995
Saab 9-5 Arc	3	36 955
Volvo S60 2.5T	3	33 890

a. Trace un diagrama de dispersión con estos datos tomando como variable independiente las evaluaciones de confiabilidad.
b. Dé la ecuación de regresión obtenida por el método de mínimos cuadrados.
c. De acuerdo con este análisis, ¿cree usted que los automóviles más confiables sean más caros?
d. Estime el precio de un automóvil sedán cuya evaluación de confiabilidad sea 4.

8. Las bicicletas de montaña que actualmente cuestan menos de $1000 tienen muchos de los componentes de alta calidad que hasta hace poco sólo tenían los modelos de alta calidad. Hoy, incluso modelos de menos de $1000 suelen ofrecer suspensión flexible, pedales clipless y cuadro muy bien diseñado. Una cuestión interesante es si precios más altos corresponden a mayor facilidad de manejo, medida a través del agarre lateral de la bicicleta. Para medir el agarre lateral, *Outside Magazine* empleó una escala de evaluación del 1 al 5, en la que el 1 correspondía a mala y 5 a promedio. A continuación se presenta el agarre lateral y los precios de 10 bicicletas de montaña probadas por *Outside Magazine* (*Outside Magazine Buyer's Guide,* 2001)



Fabricante y modelo	Agarre lateral	Precio ($)
Raleigh M80	1	600
Marin Bear Valley Feminina	1	649
GT Avalanche 2.0	2	799
Kona Jake the Snake	1	899
Schwinn Moab 2	3	950
Giant XTC NRS 3	4	1100
Fisher Paragon Genesisters	4	1149
Jamis Dakota XC	3	1300
Trek Fuel 90	5	1550
Specialized Stumpjumper M4	4	1625

a. Trace un diagrama de dispersión con estos datos tomando como variable independiente el agarre lateral.
b. ¿Parecen indicar estos datos que los modelos más caros sean de más fácil manejo? Explique.
c. Dé la ecuación de regresión estimada obtenida por el método de mínimos cuadrados.
d. ¿Cuál es el precio estimado de una bicicleta de montaña cuyo agarre lateral tenga una evaluación de 4?

9. Un gerente de ventas recolectó los datos siguientes sobre ventas anuales y años de experiencia.



Vendedor	Años de experiencia	Ventas anuales (miles de $)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

a. Elabore un diagrama de dispersión con estos datos, en el que la variable independiente sean los años de experiencia.
b. Dé la ecuación de regresión estimada que puede emplearse para predecir las ventas anuales cuando se conocen los años de experiencia.
c. Use la ecuación de regresión estimada para pronosticar las ventas anuales de un vendedor de 9 años de experiencia.

10. Bergans of Norway ha estado fabricando equipo para excursionismo desde 1908. En los datos que se presentan en la tabla siguiente se da la temperatura (°F) y el precio ($) de 11 modelos de sacos de dormir fabricados por Bergans (*Backpacker* 2006 Gear Guide)



Modelo	Temperatura	Precio
Ranger 3-Seasons	12	319
Ranger Spring	24	289
Ranger Winter	3	389
Rondane 3-Seasons	13	239
Rondane Summer	38	149
Rondane Winter	4	289
Senja Ice	5	359
Senja Snow	15	259
Senja Zero	25	229
Super Light	45	129
Tight & Light	25	199

a. Trace un diagrama de dispersión con estos datos, en el que la variable independiente sea la temperatura (°F).
b. ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre temperatura y precio?
c. Use el método de mínimos cuadrados para obtener la ecuación de regresión estimada.
d. Prediga cuál será el precio de un saco de dormir si el índice de temperatura (°F) es 20.

11. Aunque actualmente en los aeropuertos grandes los retrasos son menos frecuentes, es útil saber en qué aeropuertos es más probable que le echen a perder a uno sus planes. Además, si su vuelo llega con retraso a un determinado aeropuerto en el que tiene que hacer un trasbordo, ¿cuál es la probabilidad de que se retrase la salida y que pueda hacer así el trasbordo? En la tabla siguiente se muestra el porcentaje de llegadas y salidas retrasadas durante el mes de agosto en 13 aeropuertos (*Business 2.0,* febrero 2002).



Aeropuerto	Llegadas retrasadas (%)	Salidas retrasadas (%)
Atlanta	24	22
Charlotte	20	20
Chicago	30	29
Cincinnati	20	19
Dallas	20	22
Denver	23	23
Detroit	18	19
Houston	20	16
Minneapolis	18	18
Phoenix	21	22
Pittsburgh	25	22
Salt Lake City	18	17
St. Louis	16	16

a. Trace un diagrama de dispersión con estos datos, en el que la variable independiente sean las llegadas retrasadas.
b. ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre llegadas retrasadas y salidas retrasadas?
c. Use el método de mínimos cuadrados para obtener la ecuación de regresión estimada.
d. ¿Cómo se debe interpretar la pendiente de la ecuación de regresión estimada?
e. Suponga que en el aeropuerto de Filadelfia hubo 22% de llegadas retrasadas. ¿Cuál es el porcentaje estimado de salidas retrasadas?

12. Una moto acuática personal (personal watercraft, PWC) es una embarcación a motor dentro de borda diseñada para ser conducida por una persona sentada, de pie o arrodillada. Al principio de los años 80, Kawasaki Motors Corp. USA introdujo la moto acuática JET SKI©, la primera moto acuática comercial. Hoy *jet ski* se usa como término genérico para motos acuáticas personales. En la tabla siguiente se dan pesos (redondeados a la decena de libra más cercana) y precios (redondeados a los 50 dólares más cercanos) de 10 motos acuáticas personales de tres plazas (www.jetskinews.com, 2006).



Fabricante y modelo	Peso (lb)	Precio ($)
Honda AquaTrax F-12	750	9 500
Honda AquaTrax F-12X	790	10 500
Honda AquaTrax F-12X GPScape	800	11 200
Kawasaki STX-12F Jetski	740	8 500
Yamaha FX Cruiser Waverunner	830	10 000
Yamaha FX High Output Waverunner	770	10 000
Yamaha FX Waverunner	830	9 300
Yamaha VX110 Deluxe Waverunner	720	7 700
Yamaha VX110 Sport Waverunner	720	7 000
Yamaha XLT1200 Waverunner	780	8 500

a. Trace el diagrama de dispersión correspondiente a estos datos, empleando el peso como variable independiente.
b. ¿Qué indica el diagrama de dispersión del inciso a) respecto a la relación entre peso y precio?
c. Use el método de mínimos cuadrados para obtener la ecuación de regresión estimada.
d. Indique cuál será el precio de una moto acuática de tres plazas cuyo peso sea 750 libras.
e. La Honda Aqua Trax F-12 pesa 750 libras y su precio es $9500. ¿No debería ser el precio pronosticado en el inciso d) también de $950?

f. La Jetski Kawasaki SX-R 800 tiene asiento para una persona y pesa 350 libras. ¿Cree usted que la ecuación de regresión estimada obtenida en el inciso c) deba emplearse para predecir su precio?

13. Para la Dirección general de impuestos internos de Estados Unidos el que las deducciones parezcan razonables depende del ingreso bruto ajustado del contribuyente. Deducciones grandes que comprenden deducciones por donaciones de caridad o por atención médica son más probables en contribuyentes que tengan un ingreso bruto ajustado grande. Si las deducciones de un contribuyente son mayores que las correspondientes a un determinado nivel de ingresos, aumentan las posibilidades de que se le realice una auditoría.

Ingreso bruto ajustado (miles de $)	**Monto razonable de las deducciones (miles de $)**
22	9.6
27	9.6
32	10.1
48	11.1
65	13.5
85	17.7
120	25.5

a. Trace un diagrama de dispersión con estos datos empleando como variable independiente el ingreso bruto ajustado.
b. Use el método de mínimos cuadrados para obtener la ecuación de regresión estimada
c. Si el ingreso bruto ajustado de un contribuyente es $52 500, estime el monto razonable de deducciones. Si el contribuyente tiene deducciones por $20 400, ¿estará justificada una auditoría? Explique.

14. Los salarios iniciales de contadores y auditores en Rochester, Nueva York, corresponden a los de muchos ciudadanos de Estados Unidos. En la tabla siguiente se presentan salarios iniciales (en miles de dólares) y el índice del costo de vida en Rochester y en otras nueve zonas metropolitanas (*Democrat and Chronicle*, 1 de septiembre de 2002).



Salaries

Área metropolitana	**Índice**	**Salario (miles de $)**
Oklahoma City	82.44	23.9
Tampa/St. Petersburg/Clearwater	79.89	24.5
Indianapolis	55.53	27.4
Buffalo/Niagara Falls	41.36	27.7
Atlanta	39.38	27.1
Rochester	28.05	25.6
Sacramento	25.50	28.7
Raleigh/Durham/Chapel Hill	13.32	26.7
San Diego	3.12	27.8
Honolulu	0.57	28.3

a. Elabore un diagrama de dispersión con estos datos empleando como variable independiente el índice del costo de vida.
b. Obtenga la ecuación de regresión para relacionar el índice del costo de vida con el salario inicial.
c. Estime el salario inicial en una zona metropolitana en la que el índice del costo de vida es 50.



Coeficiente de determinación

En el ejemplo de Armand Pizza Parlors para aproximar la relación lineal entre el tamaño de la población de estudiantes x y las ventas trimestrales y se obtuvo la ecuación de regresión estimada $\hat{y} = 60 + 5x$. Ahora la pregunta es: ¿qué tan bien se ajusta a los datos la ecuación de regresión estimada? En esta sección se muestra que una medida de la bondad de ajuste de la ecuación de regresión estimada (lo bien que se ajusta la ecuación a los datos) es el **coeficiente de determinación**.

A la diferencia que existe, en la observación i, entre el valor observado de la variable dependiente y_i, y el valor estimado de la variable dependiente $\hat{y}_i$, se le llama **residual *i***. El residual i representa el error que existe al usar $\hat{y}_i$ para estimar y_i. Por lo tanto, para la observación i, el residual es $y_i - \hat{y}_i$. La suma de los cuadrados de estos residuales o errores es la cantidad que se minimiza empleando el método de los mínimos cuadrados. Esta cantidad, también conocida como *suma de cuadrados debida al error,* se denota por SCE.

SUMA DE CUADRADOS DEBIDA AL ERROR

$$\text{SCE} = \Sigma(y_i - \hat{y}_i)^2 \quad \textbf{(14.8)}$$

El valor de SCE es una medida del error al utilizar la ecuación de regresión estimada para estimar los valores de la variable dependiente en los elementos de la muestra.

En la tabla 14.3 se muestran los cálculos que se requieren para calcular la suma de cuadrados debida al error en el ejemplo de Armand´s Pizza Parlors. Por ejemplo, los valores de las variables independiente y dependiente para/del restaurante 1 son $x_1 = 2$ y $y_1 = 58$. El valor estimado para las ventas trimestrales del restaurante 1 obtenido con la ecuación de regresión estimada es $\hat{y}_1 = 60 + 5(2) = 70$. Por lo tanto, para el restaurante 1, el error al usar $\hat{y}_1$ para estimar y_1 es $y_1 - \hat{y}_1 = 58 - 70 = -12$. El error elevado al cuadrado, $(-12)^2 = 144$, aparece en la última columna de la tabla 14.3. Después de calcular y elevar al cuadrado los residuales de cada uno de los restaurantes de la muestra, se suman y se obtiene que SCE = 1530. Por lo tanto, SCE = 1530 mide el error que existe al utilizar la ecuación de regresión estimada $\hat{y} = 60 + 5x$ para predecir las ventas.

Ahora supóngase que se pide una estimación de las ventas trimestrales sin saber cuál es el tamaño de la población de estudiantes. Sin tener conocimiento de ninguna otra variable relacionada con las ventas trimestrales, se emplearía la media muestral como una estimación de las ven-

TABLA 14.3 CÁLCULO DE SCE EN EL EJEMPLO ARMAND’S PIZZA PARLORS

Restaurante i	x_i = población de estudiantes (miles)	y_i = ventas trimestrales (miles de $)	Ventas pronosticadas $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Error al cuadrado $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SCE = 1530

TABLA 14.4 CÁLCULO DE LA SUMA TOTAL DE CUADRADOS EN EL EJEMPLO DE ARMAND'S PIZZA PARLORS

Restaurante i	x_i = población de estudiantes (miles)	y_i = ventas trimestrales (miles de $)	Desviación $y_i - \bar{y}$	Desviación al cuadrado $(y_i - \bar{y})^2$
1	2	58	−72	5 184
2	6	105	−25	625
3	8	88	−42	1 764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1 521
9	22	149	19	361
10	26	202	72	5 184
				STC = 15 730

tas trimestrales de/en cualquiera de los restaurantes. En la tabla 14.2 se mostró que de acuerdo con los datos de las ventas, $\Sigma y_i = 1300$. Por lo tanto, la media de las ventas trimestrales en la muestra de los 10 restaurantes Armand's es $\bar{y} = \Sigma y_i/n = 1300/10 = 130$. En la tabla 14.4 se presenta la suma de las desviaciones al cuadrado que se obtiene cuando se usa la media muestral $\bar{y} = 130$ para estimar el valor de las ventas trimestrales de cada uno de los restaurantes de la muestra. Para el i-ésimo restaurante de la muestra, la diferencia $y_i - \bar{y}$ proporciona una medida del error que hay al usar $\bar{y}$ para estimar las ventas. La correspondiente suma de cuadrados, llamada *suma total de cuadrados*, se denota STC.

SUMA TOTAL DE CUADRADOS

$$STC = \Sigma(y_i - \bar{y})^2 \quad \textbf{(14.9)}$$

La suma debajo de la última columna de la tabla 14.4 es la suma total de cuadrados en el ejemplo de Armand's Pizza Parlors; esta suma es STC = 15 730.

Como STC = 15 730 y SCE = 1530, la línea de regresión estimada se ajusta mucho mejor a los datos que la línea $y = \bar{y}$.

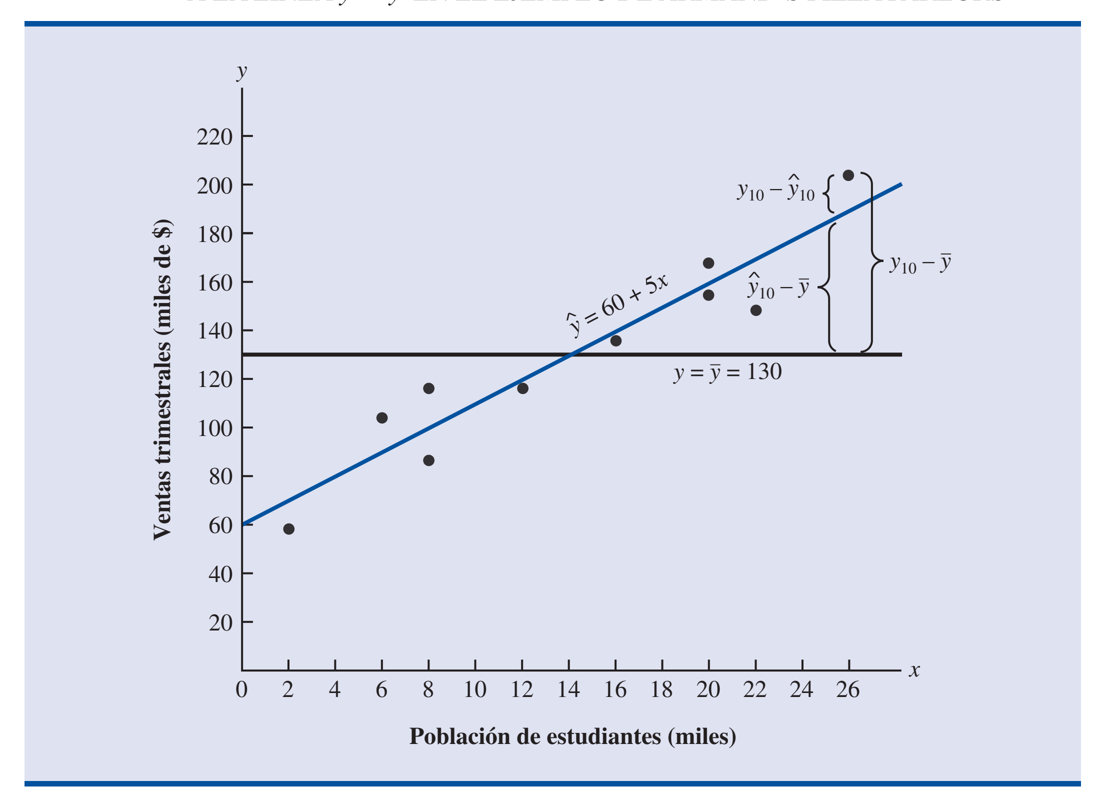
En la figura 14.5 se muestra la línea de regresión estimada $\hat{y} = 60 + 5x$ y la línea correspondientes a $\bar{y} = 130$. Obsérvese que los puntos se encuentran más agrupados en torno a la recta de regresión estimada que en torno a la línea $\bar{y} = 130$. Por ejemplo, se ve que para el 10o. restaurante de la muestra, el error es mucho más grande cuando se usa $\bar{y} = 130$ para estimar y_{10} que cuando se usa $\hat{y}_{10} = 60 + 5(26) = 190$. Se puede entender STC como una medida de qué tanto se agrupan las observaciones en torno a la recta $\bar{y}$ y SCE como una medida de qué tanto se agrupan las observaciones en torno de la recta $\hat{y}$.

Para medir qué tanto se desvían de $\bar{y}$ los valores $\bar{y}$, de la recta de regresión, se calcula otra suma de cuadrados. A esta suma se le llama *suma de cuadrados debida a la regresión* y se denota SCR.

SUMA DE CUADRADOS DEBIDA A LA REGRESIÓN

$$SCR = \Sigma(\hat{y}_i - \bar{y})^2 \quad \textbf{(14.10)}$$

FIGURA 14.5 DESVIACIONES RESPECTO A LA LÍNEA DE REGRESIÓN ESTIMADA Y A LA LÍNEA $y = \bar{y}$ EN EL EJEMPLO DE ARMAND'S PIZZA PARLORS



Por lo antes dicho, se esperaría que hubiera alguna relación entre STC, SCR y SCE. En efecto, y la relación entre estas tres sumas de cuadrados constituye uno de los resultados más importantes de la estadística.

La SCR puede entenderse como la parte explicada de la STC, y la SCE puede entenderse como la parte no explicada de la STC.

RELACIÓN ENTRE STC, SCR Y SCE

$$STC = SCR + SCE \quad \textbf{(14.11)}$$

donde

STC = suma total de cuadrados
SCR = suma de cuadrados debida a la regresión
SCE = suma de cuadrados debida al error

La ecuación (14.11) muestra que la suma total de cuadrados puede ser dividida en dos componentes, la suma de los cuadrados debida a la regresión y la suma de cuadrados debida al error. Por lo tanto, si se conocen dos cualesquiera de estas sumas de cuadrados, es fácil calcular la tercera suma de cuadrados. Por ejemplo, en el ejemplo de Armand's Pizza Parlors, se conocen SCE = 1530 y STC 15 730; por lo tanto, despejando de la ecuación (14.11) SCR, se encuentra que la suma de los cuadrados debidos a la regresión es

$$SCR = STC + SCE = 15\,730 - 1530 = 14\,200$$

Ahora se verá cómo se usan estas tres sumas de cuadrados, STC, SCR y SCE, para obtener una medida de la bondad de ajuste de la ecuación de regresión estimada. La ecuación de regresión estimada se ajustaría perfectamente a los datos si cada uno de los valores de la variable independiente y_i se encontraran sobre la recta de regresión. En este caso para todas las observaciones se tendría que $y_i - \hat{y}_i$ sería igual a cero, con lo que SCE $= 0$. Como STC $=$ SCR $+$ SCE se ve que para que haya un ajuste perfecto SCR debe ser igual a STC, y el cociente (SCR/STC) debe ser igual a uno. Cuando los ajustes son malos, se tendrán valores altos para SCE. Si en la ecuación (14.11) se despeja SCE, se tiene que SCE $=$ STC $-$ SCR. Por lo tanto, los valores más grandes de SCE (y por lo tanto un peor ajuste) se presentan cuando SCR $= 0$ y SCE $=$ STC.

El cociente SCR/STC, que toma valores entre cero y uno, se usa para evaluar la bondad de ajuste de la ecuación de regresión estimada. A este cociente se le llama *coeficiente de determinación* y se denota r^2.

COEFICIENTE DE DETERMINACIÓN

$$r^2 = \frac{\text{SCR}}{\text{STC}} \tag{14.12}$$

En el ejemplo de Armand's Pizza Parlors, el valor del coeficiente de determinación es

$$r^2 = \frac{\text{SCR}}{\text{STC}} = \frac{14\,200}{15\,730} = 0.9027$$

Si se expresa el coeficiente de determinación en forma de porcentaje, r^2 se puede interpretar como el porcentaje de la suma total de cuadrados que se explica mediante el uso de la ecuación de regresión estimada. En el ejemplo de Armand's Pizza Parlors, se concluye que 90.27% de la variabilidad en las ventas se explica por la relación lineal que existe entre el tamaño de la población de estudiantes y las ventas. Sería bueno que la ecuación de regresión tuviera un ajuste tan bueno.

Coeficiente de correlación

En el capítulo 3 se presentó el **coeficiente de correlación** como una medida descriptiva de la intensidad de la relación lineal entre dos variables x y y. Los valores del coeficiente de correlación son valores que van desde -1 hasta $+1$. El valor $+1$ indica que las dos variables x y y están perfectamente relacionadas en una relación lineal positiva. Es decir, los puntos de todos los datos se encuentran en una línea recta que tiene pendiente positiva. El valor -1 indica que x y y están perfectamente relacionadas, en una relación lineal negativa, todos los datos se encuentran en una línea recta que tiene pendiente negativa. Los valores del coeficiente de correlación cercanos a cero indican que x y y no están relacionadas linealmente.

En la sección 3.5 se presentó la ecuación para calcular el coeficiente de correlación muestral. Cuando se ha realizado un análisis de regresión y se ha calculado el coeficiente de determinación r^2, el coeficiente de correlación muestral se puede calcular como se indica a continuación.

COEFICIENTE DE CORRELACIÓN MUESTRAL

$$\begin{aligned} r_{xy} &= (\text{signo de } b_1)\sqrt{\text{Coeficiente de determinación}} \\ &= (\text{signo de } b_1)\sqrt{r^2} \end{aligned} \tag{14.13}$$

donde

$$b_1 = \text{pendiente de la ecuación de regresión estimada } \hat{y} = b_0 + b_1x$$

El signo del coeficiente de regresión muestral es positivo si la ecuación de regresión tiene pendiente positiva ($b_1 > 0$) y es negativo si la ecuación de regresión estimada tiene pendiente negativa ($b_1 < 0$).

En el ejemplo de Armand's Pizza Parlor, el valor del coeficiente de determinación correspondiente a la ecuación de regresión estimada $\hat{y} = 60 + 5x$ es 0.9027. Como la pendiente de la ecuación de regresión estimada es positiva, la ecuación (14.13) indica que el coeficiente de correlación muestral es $+\sqrt{0.9027} = +0.9501$ Con este coeficiente de correlación muestral, $r_{xy} = +0.9501$, se concluye que existe una relación lineal fuerte entre x y y.

En el caso de una relación lineal entre dos variables, tanto el coeficiente de determinación como el coeficiente de correlación muestral proporcionan medidas de la intensidad de la relación. El coeficiente de determinación proporciona una medida cuyo valor va desde cero hasta uno, mientras que el coeficiente de correlación muestral proporciona una medida cuyo valor va desde –1 hasta +1. El coeficiente de correlación lineal está restringido a la relación lineal entre dos variables, pero el coeficiente de determinación puede emplearse para relaciones no lineales y para relaciones en las que hay dos o más variables independientes. Por tanto, el coeficiente de determinación tiene un rango más amplio de aplicaciones.

NOTAS Y COMENTARIOS

1. Al obtener la ecuación de regresión estimada mediante el método de mínimos cuadrados y calcular el coeficiente de determinación, no se hizo ninguna suposición probabilística acerca del término del error ϵ ni tampoco una prueba de significancia para la relación entre x y y. Los valores grandes de r^2 implican que la recta de mínimos cuadrados se ajusta mejor a los datos; es decir, las observaciones se encuentran más cerca de la recta de mínimos cuadrados. Sin embargo, usando únicamente r^2 no se pueden sacar conclusiones acerca de si la relación entre x y y es estadísticamente significativa. Tal conclusión debe basarse en consideraciones que implican el tamaño de la muestra y las propiedades de la distribución muestral adecuada de los estimadores de mínimos cuadrados.
2. Para fines prácticos, cuando se trata de datos que se encuentran en las ciencias sociales, valores de r^2 tan pequeños como 0.25 suelen considerarse útiles. En datos de la física o de las ciencias de la vida, suelen encontrarse valores de r^2 de 0.60 o mayores; en algunos casos pueden encontrarse valores mayores de 0.90. En las aplicaciones a los negocios, los valores de r^2 varían enormemente dependiendo de las características particulares de cada aplicación.

Ejercicios

Método



15. Los datos a continuación son los datos del ejercicio 1.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

La ecuación de regresión estimada para estos datos es $\hat{y} = 0.20 + 2.60x$.

a. Empleando las ecuaciones (14.8), (14.9) y (14.10) calcule SCE, STC y SCR.
b. Calcule el coeficiente de determinación r^2. Haga un comentario sobre la bondad del ajuste.
c. Calcule el coeficiente de correlación muestral.

16. Los datos a continuación son los datos del ejercicio 2.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

La ecuación de regresión estimada para estos datos es $\hat{y} = 68 - 3x$.

a. Calcule SCE, STC y SCR.
b. Calcule el coeficiente de determinación r^2. Haga un comentario sobre la bondad del ajuste.
c. Calcule el coeficiente de correlación muestral.

17. Los datos a continuación son los datos del ejercicio 3.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

La ecuación de regresión estimada para estos datos es $\hat{y} = 7.6 + 0.9x$. ¿Qué porcentaje de la suma total de cuadrados puede explicarse mediante la ecuación de regresión estimada? ¿Cuál es el valor del coeficiente de correlación muestral?

Aplicaciones



18. En los datos siguientes, y corresponde a los salarios mensuales y x es el promedio obtenido por los estudiantes que terminaron la licenciatura de administración con especialidad en sistemas de información. La ecuación de regresión estimada obtenida con estos datos es $\hat{y} = 1790.5 + 581.1x$.

Promedio	Salario mensual ($)
2.6	3300
3.4	3600
3.6	4000
3.2	3500
3.5	3900
2.9	3600

a. Calcule SCE, STC y SCR.
b. Calcule el coeficiente de determinación r^2. Haga un comentario sobre la bondad del ajuste.
c. Calcule el coeficiente de correlación muestral.

19. Los datos a continuación son los datos del ejercicio 7.



Cars

Fabricante y modelo	x = confiabilidad	y = precio ($)
Acura TL	4	33 150
BMW 330i	3	40 570
Lexus IS300	5	35 105
Lexus ES330	5	35 174
Mercedes-Benz C320	1	42 230
Lincoln LS Premium (V6)	3	38 225
Audi A4 3.0 Quattro	2	37 605
Cadillac CTS	1	37 695
Nissan Maxima 3.5 SE	4	34 390
Infiniti I35	5	33 845
Saab 9-3 Aero	3	36 910
Infiniti G35	4	34 695
Jaguar X-Type 3.0	1	37 995
Saab 9-5 Arc	3	36 955
Volvo S60 2.5T	3	33 890

La ecuación de regresión estimada para estos datos es $\hat{y} = 40\,639 - 1301.2x$. ¿Qué porcentaje de la suma total de cuadrados puede explicarse mediante la ecuación de regresión estimada? Haga un comentario sobre la bondad del ajuste ¿Cuál es el valor del coeficiente de correlación muestral?

20. *Consumer Reports* publica pruebas y evaluaciones sobre televisores de alta definición. Para cada modelo se elaboró una evaluación general basada principalmente en la calidad de la imagen. Una evaluación más alta indica un mejor funcionamiento. En los datos siguientes se dan evaluación general y precio de televisores de plasma de 45 pulgadas (*Consumer Reports*, marzo 2006).



PlasmaTV

Marca	Precio	Puntuación en la valuación
Dell	2800	62
Hisense	2800	53
Hitachi	2700	44
JVC	3500	50
LG	3300	54
Maxent	2000	39
Panasonic	4000	66
Phillips	3000	55
Proview	2500	34
Samsung	3000	39

a. Use estos datos para obtener una ecuación de regresión estimada que pueda emplearse para estimar la puntuación en la evaluación general de una televisión de 42 pulgadas dado el precio.
b. Calcule r^2. ¿Proporcionó un buen ajuste la ecuación de regresión estimada?
c. Estime la puntuación en la evaluación general de un televisor cuyo precio es $3200.

21. Una aplicación importante del análisis de regresión a la contaduría es la estimación de costos. Con datos sobre volumen de producción y costos y empleando el método de mínimos cuadrados para obtener la ecuación de regresión estimada que relacione volumen de producción y costos, los contadores pueden estimar los costos correspondientes a un determinado volumen de producción. Considere la siguiente muestra de datos sobre volumen de producción y costos totales de una operación de fabricación.

Volumen de producción (unidades)	Costos totales ($)
400	4000
450	5000
550	5400
600	5900
700	6400
750	7000

a. Con estos datos obtenga la ecuación de regresión estimada para pronosticar los costos totales dado un volumen de producción.
b. ¿Cuál es el costo por unidad producida?
c. Calcule el coeficiente de determinación. ¿Qué porcentaje de la variación en los costos totales puede ser explicada por el volumen de producción?
d. De acuerdo con el programa de producción de la empresa, el mes próximo se deberán producir 500 unidades. ¿Cuál es el costo total estimado de esta operación?

22. *PC World* publicó evaluaciones de las cinco mejores impresoras láser de oficina y de las cinco mejores impresoras láser corporativas (*PC World*, febrero 2003). La impresora de oficina mejor evaluada fue la Minolta-QMS PagePro 1250W, que en la evaluación general obtuvo una puntuación de 91 puntos. La impresora láser corporativa mejor evaluada fue la Xerox Phase 4400/N, que

en la evaluación general obtuvo una puntuación de 83 puntos. En la tabla siguiente se da rapidez, en páginas por minuto (ppm), en la impresión de texto y precio de cada impresora.



Printers

Nombre	Tipo	Velocidad (ppm)	Precio ($)
Minolta-QMS PagePro 1250W	Oficina	12	199
Brother HL-1850	Oficina	10	499
Lexmark E320	Oficina	12.2	299
Minolta-QMS PagePro 1250E	Oficina	10.3	299
HP Laserjet 1200	Oficina	11.7	399
Xerox Phaser 4400/N	Corporativa	17.8	1850
Brother HL-2460N	Corporativa	16.1	1000
IBM Infoprint 1120n	Corporativa	11.8	1387
Lexmark W812	Corporativa	19.8	2089
Oki Data B8300n	Corporativa	28.2	2200

a. Dé la ecuación de regresión estimada empleando velocidad como variable independiente.
b. Calcule r^2. ¿Qué porcentaje de la variación del precio puede ser explicado por la velocidad de la impresora?
c. ¿Cuál es el coeficiente de correlación muestral entre velocidad y precio? ¿Refleja este coeficiente una relación fuerte o débil entre la velocidad de la impresora y el costo?



Suposiciones del modelo

En un análisis de regresión se empieza por hacer una suposición acerca del modelo apropiado para la relación entre las variables dependientes e independientes. En el caso de la regresión lineal simple, se supone que el modelo de regresión es

$$y = \beta_0 + \beta_1 x + \epsilon$$

Después empleando el método de mínimos cuadrados se obtienen los valores de b_0 y b_1, que son las estimaciones de los parámetros β_0 y β_1, respectivamente, del modelo. Así se llega la ecuación de regresión estimada

$$\hat{y} = b_0 + b_1 x$$

Como se vio, el valor del coeficiente de determinación (r^2) es una medida de la bondad de ajuste de la ecuación de regresión estimada. Sin embargo, aun cuando se obtenga un valor grande para r^2, la ecuación de regresión estimada no debe ser usada hasta que se realice un análisis para determinar si el modelo empleado es adecuado. Un paso importante para ver si el modelo empleado es adecuado es probar la significancia de la relación. Las pruebas de significancia en el análisis de regresión están basadas en las suposiciones siguientes acerca del término del error ϵ.

SUPOSICIONES ACERCA DEL TÉRMINO DEL ERROR EN EL ANÁLISIS DE REGRESIÓN

$$y = \beta_0 + \beta_1 x + \epsilon$$

1. El término del error ϵ es una variable aleatoria cuya media, o valor esperado, es cero; es decir, $E(\epsilon) = 0$.
Implicación: β_0 y β_1 son constantes, por lo tanto $E(\beta_0) = \beta_0$ y $E(\beta_1) = \beta_1$; así, para un valor dado de x, el valor esperado de y es

$$E(y) = \beta_0 + \beta_1 x \qquad \textbf{(14.14)}$$

(*continúa*)

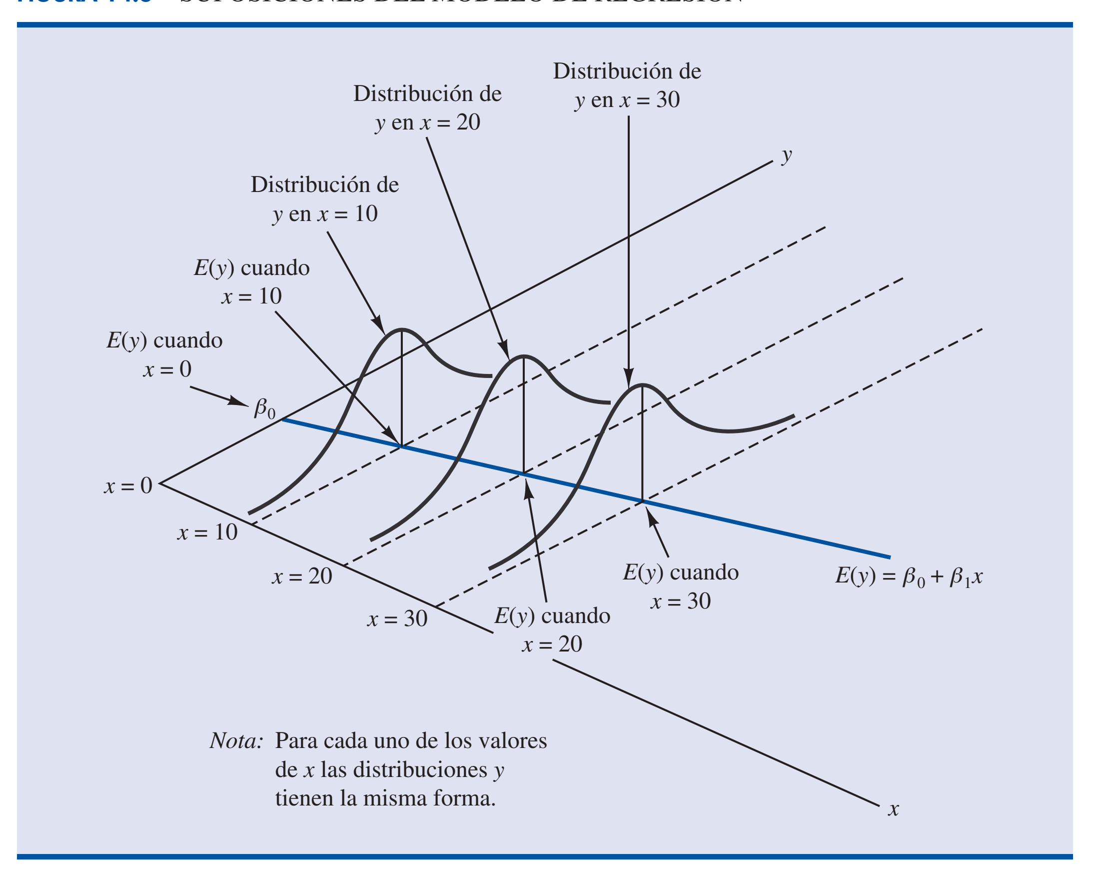
Como ya se indicó, a la ecuación (14.14) se le conoce como ecuación de regresión.

2. La varianza de ϵ, que se denota σ^2, es la misma para todos los valores de x.
   Implicación: La varianza de y respecto a la recta de regresión es igual a σ^2 y es la misma para todos los valores de x.
3. Los valores de ϵ son independientes.
   Implicación: El valor de ϵ correspondiente a un determinado valor de x no está relacionado con el valor de ϵ correspondiente a ningún otro valor de x; por lo tanto, el valor de y correspondiente a un determinado valor de x no está relacionado con el valor de y de ningún otro valor de x.
4. El término del error ϵ es una variable aleatoria distribuida normalmente.
   Implicación: como y es función lineal de ϵ, también y es una variable aleatoria distribuida normalmente.

En la figura 14.6 se muestran las suposiciones del modelo y sus implicaciones; obsérvese que en esta interpretación gráfica, el valor de $E(y)$ cambia de acuerdo con el valor de x que se considere. Sin embargo, sea cual sea el valor de x, la distribución de probabilidad de ϵ, y por tanto la distribución de probabilidad de y, son distribuciones normales, que tienen, todas, la misma varianza. El valor específico del error ϵ en cualquier punto depende de si el valor real de y es mayor o menor que $E(y)$.

En este punto, hay que tener presente que también se hace una suposición o se tienen una hipótesis acerca de la forma de la relación entre x y y. Es decir, se supone que la base de la relación entre las variables es una recta representada por $\beta_0 + \beta_1 x$. No se debe perder de vista el

FIGURA 14.6 SUPOSICIONES DEL MODELO DE REGRESIÓN



hecho de que puede haber algún otro modelo, por ejemplo $y = \beta_0 + \beta_1 x^2 + \epsilon$, que resulte ser un mejor modelo para la relación en estudio.



14.5 Prueba de significancia

En una ecuación de regresión lineal simple, la media o valor esperado de y es una función lineal de x: $E(y) = \beta_0 + \beta_1 x$. Pero si el valor de β_1 es cero, $E(y) = \beta_0 + (0)x = \beta_0$. En este caso, el valor medio de y no depende del valor de x y por lo tanto se puede concluir que x y y no están relacionadas linealmente. Pero si el valor de β_1 es distinto de cero, se concluirá que las dos variables están relacionadas. Por lo tanto, para probar si existe una relación de regresión significante, se debe realizar una prueba de hipótesis para determinar si el valor de β_1 es distinto de cero. Hay dos pruebas que son las más usadas. En ambas, se requiere una estimación de σ^2, la varianza de ϵ en el modelo de regresión.

Estimación de σ^2

De acuerdo con el modelo de regresión y con sus suposiciones, se puede concluir que σ^2, la varianza de ϵ, representa también la varianza de los valores de y respecto a la recta de regresión. Recuérdese que a las desviaciones de los valores de y de la recta de regresión estimada se les conoce como residuales. Por lo tanto, SCE, la suma de los cuadrados de los residuales, es una medida de la variabilidad de las observaciones reales respecto a la línea de regresión estimada. El **error cuadrado medio** (ECM) proporciona una estimación de σ^2; esta estimación es SCE dividida entre sus grados de libertad.

Como $\hat{y}_i = b_0 + b_1 x_i$, SCE se puede expresar como

$$\text{SCE} = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - b_0 - b_1 x_i)^2$$

A cada suma de cuadrados le corresponde un número llamado sus grados de libertad. Se ha demostrado que SCE tiene $n - 2$ grados de libertad porque para calcular SCE es necesario estimar dos parámetros (β_0 y β_1). Por lo tanto, el cuadrado medio se calcula dividiendo SCE entre $n - 2$. ECM proporciona un estimador insesgado de σ^2. Como el valor del ECM proporciona un estimado de σ^2, se emplea también la notación s^2.

ERROR CUADRADO MEDIO (ESTIMACIÓN DE σ^2)

$$s^2 = \text{ECM} = \frac{\text{SCE}}{n - 2} \qquad \textbf{(14.15)}$$

En la sección 14.3 se encontró que en el ejemplo de Armand's Pizza Parlors, SCE = 1530; por lo tanto,

$$s^2 = \text{ECM} = \frac{1530}{8} = 191.25$$

es un estimador insesgado de σ^2.

Para estimar σ se saca la raíz cuadrada de s^2. Al valor que se obtiene, s, se le conoce como el **error estándar de estimación**.

ERROR ESTÁNDAR DE ESTIMACIÓN

$$s = \sqrt{\text{ECM}} = \sqrt{\frac{\text{SCE}}{n - 2}} \qquad \textbf{(14.16)}$$

En el ejemplo de Armand's Pizza Parlors, $s = \sqrt{\text{ECM}} = \sqrt{191.25} = 13.829$. El error estándar de estimación se emplea en la discusión siguiente acerca de las pruebas de significancia de la relación entre x y y.

Prueba *t*

El modelo de regresión lineal simple es $y = \beta_0 + \beta_1 x + \epsilon$. Si x y y están relacionadas linealmente, entonces $\beta_1 \neq 0$. El objetivo de la prueba t es determinar si se puede concluir que $\beta_1 \neq 0$. Para probar la hipótesis siguiente acerca del parámetro β_1 se emplearán los datos muestrales.

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

Si se rechaza H_0, se concluirá que $\beta_1 \neq 0$ y que entre las dos variables existe una relación estadísticamente significante. La base para esta prueba de hipótesis la proporcionan las propiedades de la distribución muestral de b_1, el estimador de β_1, obtenido mediante el método de mínimos cuadrados.

Primero, considérese que es lo que ocurriría si para el mismo estudio de regresión se usara otra muestra aleatoria simple. Supóngase, por ejemplo, que Armand's Pizza Parlors usa una muestra de las ventas de otros 10 restaurantes. El análisis de regresión de esta otra muestra dará como resultado una ecuación de regresión parecida a la ecuación de regresión anterior $\hat{y} = 60 + 5x$. Sin embargo, no puede esperarse que se obtenga exactamente la misma ecuación (una ecuación en la que la intersección con el eje y sea exactamente 60 y la pendiente sea exactamente 5). Los estimadores b_0 y b_1, obtenidos por el método de mínimos cuadrados, son estadísticos muestrales que tienen su propia distribución muestral. A continuación se presentan las propiedades de la distribución muestral de b_1.

DISTRIBUCIÓN MUESTRAL DE b_1

Valor esperado

$$E(b_1) = \beta_1$$

Desviación estándar

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} \tag{14.17}$$

Forma de la distribución

Normal

Obsérvese que el valor esperado de b_1 es β_1, por lo que b_1 es un estimador insesgado de β_1.

Como no se conoce el valor de σ, se obtiene una estimación de σ_{b_1}, que se denota s_{b_1}, estimando σ mediante s en la ecuación (14.17). De esta manera se obtiene el estimador siguiente de σ_{b_1}.

A la desviación estándar de b_1, se le conoce también como error estándar de b_1. Por lo tanto, s_{b_1} proporciona una estimación del error estándar de b_1.

DESVIACIÓN ESTÁNDAR ESTIMADA DE b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \tag{14.18}$$

En el ejemplo de Armand's Pizza Parlors, $s = 13.829$. Por lo tanto, dado que $\Sigma(x_i - \bar{x})^2 = 568$ como se muestra en la tabla 14.2, se tiene que

$$s_{b_1} = \frac{13.829}{\sqrt{568}} = 0.5803$$

es la desviación estándar estimada de b_1.

La prueba t para determinar si la relación es significativa se basa en el hecho de que el estadístico de prueba

$$\frac{b_1 - \beta_1}{s_{b_1}}$$

sigue una distribución t con $n - 2$ grados de libertad. Si la hipótesis nula es verdadera, entonces $\beta_1 = 0$ y $t = b_1/s_{b_1}$.

Ahora se realizará esta prueba de significancia con los datos de Armand's Pizza Parlors, empleando como nivel de significancia $\alpha = 0.01$. El estadístico de prueba es

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{0.5803} = 8.62$$

En los apéndices 14.3 y 14.4 se muestra el uso de Minitab y de Excel para calcular el valor-p

En las tablas de la distribución t se encuentra que para $n - 2 = 10 - 2 = 8$ grados de libertad, $t = 3.355$ da un área de 0.005 en la cola superior. Por lo tanto, el área en la cola superior de la distribución t correspondiente al valor del estadístico de prueba $t = 8.62$ debe ser menor a 0.005. Como esta prueba es una prueba de dos colas, este valor se duplica y se concluye que el valor-p para $t = 8.62$ debe ser menor a $2(0.005) = 0.01$. Empleando Excel o Minitab se encuentra valor-$p = 0.000$. Dado que el valor-p es menor a $\alpha = 0.01$ se rechaza H_0 y se concluye que β_1 no es igual a cero. Esto es suficiente evidencia para concluir que existe una relación significativa entre la población de estudiantes y las ventas trimestrales. A continuación se presenta un resumen de la prueba t de significancia para la regresión lineal simple.

PRUEBA t DE SIGNIFICANCIA PARA LA REGRESIÓN LINEAL SIMPLE

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

ESTADÍSTICO DE PRUEBA

$$t = \frac{b_1}{s_{b_1}} \tag{14.19}$$

REGLA DE RECHAZO

Método del valor-p: Rechazar H_0 si valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $t \leq -t_{\alpha/2}$ o si $t \geq t_{\alpha/2}$

donde $t_{\alpha/2}$ se toma de la distribución t con $n - 2$ grados de libertad.

Intervalo de confianza para β_1

La fórmula para un intervalo de confianza para β_1 es la siguiente:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

El estimador puntual es b_1 y el margen de error es $t_{\alpha/2}s_{b_1}$. El coeficiente de confianza para este intervalo es $1 - \alpha$ y $t_{\alpha/2}$ es el valor t que proporciona un área $\alpha/2$ en la cola superior de la distribución t con $n - 2$ grados de libertad. Supóngase, por ejemplo, que en el caso de Armand's Pizza Parlors se desea obtener una estimación de β_1 mediante un intervalo de 99% de confianza. En la tabla 2 del apéndice B se encuentra que el valor t correspondiente a $\alpha = 0.01$ y $n - 2 = 10 - 2 = 8$ grados de libertad es $t_{0.005} = 3.355$. Por lo tanto, la estimación mediante un intervalo de 99% de confianza es

$$b_1 \pm t_{\alpha/2}s_{b_1} = 5 \pm 3.355(0.5803) = 5 \pm 1.95$$

o el intervalo que va de 3.05 a 6.95.

Al emplear la prueba t de significancia la hipótesis probada fue

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

Empleando $\alpha = 0.01$ como nivel de significancia, se puede usar el intervalo de 99% de confianza como alternativa para llegar a la conclusión de la prueba de hipótesis que se obtiene con los datos de Armand's. Como 0, que es el valor hipotético de β_1, no está comprendido en el intervalo de confianza (3.05 a 6.95), se rechaza H_0 y se concluye que entre el tamaño de la población de estudiantes y las ventas trimestrales sí existe una relación estadísticamente significativa. En general, se puede usar un intervalo de confianza para probar cualquier hipótesis de dos colas acerca de β_1. Si el valor hipotético de β_1 está contenido en el intervalo de confianza, no se rechaza H_0. De lo contrario, se rechaza H_0.

Prueba *F*

Una prueba F, basada en la distribución de probabilidad F puede emplearse también para probar la significancia en la regresión. Cuando sólo se tiene una variable independiente, la prueba F lleva a la misma conclusión que la prueba t; es decir, si la prueba t indica que $\beta_1 \neq 0$ y por lo tanto que existe una relación significante, la prueba F también indicará que existe una relación significante. Pero cuando hay más de una variable independiente, sólo la prueba F puede usarse para probar que existe una relación significante general.

La lógica detrás del uso de la prueba F para determinar si la relación de regresión es estadísticamente significativa se basa en la obtención de dos estimaciones independiente de σ^2. Se explicó cómo ECM proporciona una estimación de σ^2. Si la hipótesis nula H_0: $\beta_1 = 0$ es verdadera, la suma de cuadrados debida a la regresión, SCR, dividida entre sus grados de libertad proporciona otra estimación independiente de σ^2. A esta estimación se le llama el *cuadrado medio debido a la regresión* o simplemente el *cuadrado medio de la regresión.* Y se denota CMR. En general,

$$\text{CMR} = \frac{\text{SCR}}{\text{Grados de libertad de la regresión}}$$

En los modelos que se consideran en este texto, el número de grados de libertad de la regresión es siempre igual al número de variables independientes en el modelo:

$$\text{CMR} = \frac{\text{SCR}}{\text{Número de variables independientes}} \quad \textbf{(14.20)}$$

Como en este capítulo sólo se consideran modelos de regresión con una sola variable independiente, se tiene CMR = SCR/1 = SCR. Por lo tanto, en el ejemplo de Armand's Pizza Parlors, CMR = SCR = 14 200.

Si la hipótesis nula es verdadera (H_0: $\beta_1 = 0$), CMR y ECM son dos estimaciones independientes de σ^2 y la distribución muestral de CMR/ECM sigue una distribución F en la que el nú-

mero de grados de libertad en el numerador es igual a uno y el número de grados de libertad en el denominador es igual a $n - 2$. Por lo tanto, si $\beta_1 = 0$ el valor de CMR/ECM deberá ser un valor cercano a uno. Pero, si la hipótesis nula es falsa, ($\beta_1 \neq 0$), CMR sobreestimará σ^2 y el valor de CMR/ECM se inflará; de esta manera valores grandes de CMR/ECM conducirán al rechazo de H_0 y a la conclusión de que la relación entre x y y es estadísticamente significante.

A continuación se realizará la prueba F en el ejemplo de Armand's Pizza Parlors. El estadístico de prueba es

$$F = \frac{\text{CMR}}{\text{ECM}} = \frac{14\,200}{191.25} = 74.25$$

En la regresión lineal simple, la prueba F y la prueba t proporcionan resultados idénticos.

En la tabla de la distribución F (tabla 4 del apéndice B) se observa que con un grado de libertad en el numerador y $n - 2 = 10 - 2 = 8$ grados de libertad en el denominador, $F = 11.26$ proporciona un área de 0.01 en la cola superior. Por lo tanto, el área en la cola superior de la distribución F que corresponde al estadístico de prueba $F = 74.25$ debe de ser menor a 0.01. Por lo tanto, se concluye que el valor-p debe de ser menor a $\alpha = 0.01$. Empleando Excel o Minitab se encuentra que valor-p = 0.000. Como el valor–p es menor a $\alpha = 0.01$, se rechaza H_0 y se concluye que entre el tamaño de la población de estudiantes y las ventas trimestrales, existe una relación significante. A continuación se presenta un resumen de la prueba F de significancia para la regresión lineal simple.

Si H_0 es falsa, ECM proporciona una estimación insesgada de σ^2 y el CMR sobreestima σ^2. Si H_0 es verdadera, tanto ECM como CMR proporcionan una estimación insesgada de σ^2; en este caso el valor de CMR/ECM es cercano a 1.

PRUEBA F DE SIGNIFICANCIA EN EL CASO DE LA REGRESIÓN LINEAL SIMPLE

$$H_0\colon \beta_1 = 0$$
$$H_a\colon \beta_1 \neq 0$$

ESTADÍSTICO DE PRUEBA

$$F = \frac{\text{CMR}}{\text{ECM}} \tag{14.21}$$

REGLA DE RECHAZO

Método del valor-p: Rechaza H_0 si valor-$p \leq \alpha$

Método del valor crítico: Rechaza H_0 si $F \geq F_\alpha$

donde F_α es un valor de la distribución F con 1 grado de libertad en el numerador y $n - 2$ grados de libertad en el denominador.

En el capítulo 13 se vio el análisis de varianza (ANOVA) y el uso de la **tabla de ANOVA** para proporcionar una visión resumida de los cálculos que se emplean en el análisis de varianza. Para resumir los cálculos de la prueba F de significancia para la regresión se emplea una tabla ANOVA similar. En la tabla 14.5 se presenta la forma general de una tabla ANOVA para la regresión lineal simple. En la tabla 14.6 se presenta la tabla ANOVA con los cálculos para la prueba F del ejemplo de Armand's Pizza Parlors. Regresión, error y total son los rótulos de las tres fuentes de variación, y SCR, SCE y STC las sumas de cuadrados correspondientes que aparecen en la columna 2. En la columna 3 aparecen los grados de libertad 1 para SCR, $n - 2$ para SCE y $n - 1$ para STC. Los valores de CMR y ECM aparecen en la columna 4. En la columna 5 aparece el valor de F = CMR/ECM, y en la columna 6 aparece el valor-p que corresponde al valor de F de la columna 5. Casi todos los resultados proporcionados por computadoras para el análisis de regresión presentan una tabla ANOVA de la prueba F de significancia.

En toda tabla para el análisis de varianza, la suma total de cuadrados es la suma de la suma de cuadrados de la regresión más la suma de cuadrados del error; además, el total de los grados de libertad es la suma de los grados de libertad de la regresión más los grados de libertad del error.

TABLA 14.5 FORMA GENERAL DE LA TABLA ANOVA PARA LA REGRESIÓN LINEAL SIMPLE

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Regresión	SCR	1	$\text{CMR} = \frac{\text{SCR}}{1}$	$F = \frac{\text{CMR}}{\text{CME}}$	
Error	SCE	$n - 2$	$\text{CME} = \frac{\text{SCE}}{n-2}$		
Total	STC	$n - 1$			

Algunas advertencias acerca de la interpretación de las pruebas de significancia

El análisis de regresión, que se usa para identificar la relación entre las variables, no puede emplearse como evidencia de una relación de causa y efecto.

Cuando se rechaza la hipótesis nula H_0: $\beta_1 = 0$, concluir que la relación que existe entre x y y es significativa no permite que se concluya que existe una relación de causa y efecto entre x y y. Que exista una relación de causa y efecto sólo puede concluirse cuando el analista pueda dar justificaciones teóricas de que en efecto la relación es causal. En el ejemplo de Armand's Pizza Parlors, se concluye que existe una relación significante entre el tamaño de la población de estudiantes x y las ventas trimestrales y; aún más, la ecuación de regresión estimada $\hat{y} = 60 + 5x$ da una estimación de la relación obtenida por el método de mínimos cuadrados. Sin embargo, por el solo hecho de que se haya encontrado que hay una relación estadísticamente significativa entre x y y, no se puede concluir que cambios en la población de estudiantes x *causen* cambios en las ventas trimestrales y. Si es posible concluir que haya una relación de causa y efecto se deja a las justificaciones teóricas y a la opinión de los analistas. Los administradores de Armand's creían que el aumento en la población de estudiantes probablemente fuera una causa del aumento de las ventas trimestrales. Por lo tanto, el resultado de la prueba de significancia les permite concluir que hay una relación de causa y efecto.

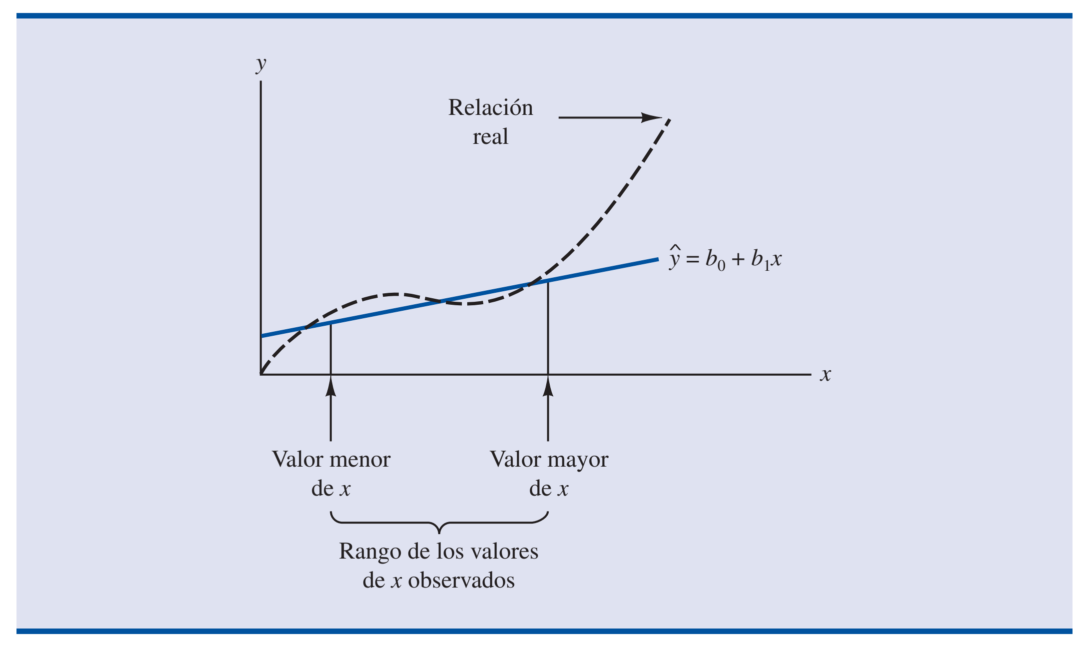
Además, el hecho de que se pueda rechazar H_0: $\beta_1 = 0$ y demostrar que hay significancia estadística no permite concluir que la relación entre x y y sea lineal. Lo único que se puede decir es que x y y están relacionadas y que la relación lineal explica una porción significante de la variabilidad de y sobre el rango de los valores de x observados en la muestra. En la figura 14.7 se ilustra esta relación. La prueba de significancia lleva al rechazo de la hipótesis nula H_0: $\beta_1 = 0$ y a la hipótesis de que x y y están significantemente relacionadas, pero en la figura se observa que la verdadera relación entre x y y no es lineal. Aunque la aproximación lineal proporcionada

TABLA 14.6 TABLA ANOVA PARA EL EJEMPLO DE ARMAND'S PIZZA PARLORS

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Regresión	14 200	1	$\frac{14\,200}{1} = 14\,200$	$\frac{14\,200}{191.25} = 74.25$	0.000
Error	1 530	8	$\frac{1530}{8} = 191.25$		
Total	15 730	9			



FIGURA 14.7 EJEMPLO DE UNA APROXIMACIÓN LINEAL PARA UNA RELACIÓN QUE NO ES LINEAL



por $\hat{y} = b_0 + b_1x$ es buena en el rango de los valores observados de x en la muestra, se vuelve deficiente fuera de ese rango.

Dada una relación significante, la ecuación de regresión estimada se puede usar con confianza para predicciones correspondientes a valores de x dentro del rango de los valores de x observados en la muestra. En el ejemplo de Armand's Pizza Parlors, este rango corresponde a los valores de x entre 2 y 26. A menos que haya otras razones que indiquen que el modelo es válido más allá de este rengo, las predicciones fuera del rango de la variable independiente deben hacerse con cuidado. En el ejemplo de Armand's Pizza Parlors, como se ha encontrado que la relación de regresión es significante al nivel de significancia de 0.01, se puede tener confianza para usar esta relación para predecir las ventas de restaurantes en los que la población de estudiantes correspondiente esté en el intervalo de 2000 a 26 000.

NOTAS Y COMENTARIOS

1. Las suposiciones hechas acerca del término del error (sección 14.4) son las que permiten las pruebas de significancia estadística de esta sección. Las propiedades de la distribución muestral de b_1 y las subsiguientes pruebas t y F siguen directamente de estas suposiciones.
2. No se debe confundir la significancia estadística con la significancia práctica. Con tamaños de muestra muy grandes, se pueden obtener resultados estadísticamente significantes para valores pequeños de b_1; en tales casos hay que tener cuidado al concluir que la relación tiene significancia práctica.
3. Una prueba de significancia para la relación lineal entre x y y también se puede realizar usando el coeficiente de correlación muestral r_{xy}. Empleando r_{xy} para denotar el coeficiente de correlación poblacional, las hipótesis son las siguientes.

$$H_0\colon \rho_{xy} = 0$$
$$H_a\colon \rho_{xy} \neq 0$$

Si se rechaza H_0, se puede concluir que existe una relación significante. En el apéndice 14.2 se proporcionan los detalles de esta prueba. Sin embargo, las pruebas t y F presentadas en esta sección dan el mismo resultado que la prueba de significancia usando el coeficiente de correlación. Por lo tanto, si ya se ha realizado una prueba t o una prueba F no es necesario realizar una prueba de significancia usando el coeficiente de correlación.

Ejercicios

Métodos



23. A continuación se presentan los datos del ejercicio 1.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Usando la ecuación (14.15) calcule el error cuadrado medio.
b. Usando la ecuación (14.16) calcule el error estándar de estimación.
c. Usando la ecuación (14.18) calcule la desviación estándar estimada de b_1.
d. Use la prueba t para probar las hipótesis siguientes ($\alpha = 0.05$)

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

e. Use la prueba F para probar las hipótesis del inciso d) empleando como nivel de significancia 0.05. Presente los resultados en el formato de tabla de análisis de varianza.

24. A continuación se presentan los datos del ejercicio 2.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a. Usando la ecuación (14.15) calcule el error cuadrado medio.
b. Usando la ecuación (14.16) calcule el error estándar de estimación.
c. Usando la ecuación (14.18) calcule la desviación estándar estimada de b_1.
d. Use la prueba t para probar las hipótesis siguientes ($\alpha = 0.05$).

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

e. Use la prueba F para probar las hipótesis del inciso d) empleando como nivel de significancia 0.05. Presente los resultados en el formato de tabla de análisis de varianza.

25. A continuación se presentan los datos del ejercicio 3.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

a. ¿Cuál es el valor del error estándar de estimación?
b. Pruebe si existe una relación significante usando la prueba t. Use $\alpha = 0.05$.
c. Emplee la prueba F para ver si existe una relación significante. Use $\alpha = 0.05$. ¿Cuál es la conclusión?

Aplicaciones



26. En el ejercicio 18 los datos sobre promedio obtenido en la licenciatura y salarios mensuales fueron los siguientes.

Promedio	Salario mensual ($)	Promedio	Salario mensual ($)
2.6	3300	3.2	3500
3.4	3600	3.5	3900
3.6	4000	2.9	3600

a. ¿Indica la prueba t que haya una relación significante entre promedio y salario mensual?
b. Pruebe si la relación es significante usando la prueba F. ¿Cuál es la conclusión? Use $\alpha = 0.05$.
c. Dé la tabla ANOVA.

27. La revista *Outside Magazine* probó 10 modelos de mochilas y botas para excursionismo. En la tabla siguiente se presentan los datos de soporte superior y precio de cada modelo. El soporte superior se midió con una escala del 1 al 5 en la que 1 significa aceptable y 5 denota excelente soporte superior (*Outside Magazine Buyer´s Guide 2001*).


Boots

Fabricante y modelo	Soporte superior	Precio ($)
Salomon Super Raid	2	120
Merrell Chameleon Prime	3	125
Teva Challenger	3	130
Vasque Fusion GTX	3	135
Boreal Maigmo	3	150
L.L. Bean GTX Super Guide	5	189
Lowa Kibo	5	190
Asolo AFX 520 GTX	4	195
Raichle Mt. Trail GTX	4	200
Scarpa Delta SL M3	5	220

a. Use estos datos para obtener la ecuación de regresión estimada para estimar el precio de las mochilas y las botas para excursionismo con base en el soporte superior.
b. Empleando $\alpha = 0.05$, determine si hay relación entre soporte superior y precio.
c. Confiaría en usar la ecuación de regresión estimada obtenida en el inciso a) para estimar el precio de las mochilas y botas para excursión con base en la evaluación del soporte superior.
d. Estime el precio de una mochila que tiene un 4 como evaluación del soporte superior

28. En el ejercicio 10, con los datos de temperatura (°F) y precio ($) de 11 sacos de dormir de Bergans de Norway se obtuvo la ecuación de regresión estimada $\hat{y} = 359.2668 - 5.2772x$. Empleando 0.05 como nivel de significancia, determine si temperatura y precio están relacionados. Dé la tabla de ANOVA. ¿Cuál es la conclusión?

29. Vuelva al ejercicio 21, en el que se usaron los datos sobre volumen de producción y costos para obtener una ecuación de regresión estimada que relacionaba el volumen de producción y los costos de una determinada operación de producción. Use $\alpha = 0.05$ para determinar si el volumen de producción está relacionado de manera significativa con los costos totales. Dé la tabla ANOVA. ¿Cuál es la conclusión?

30. Vuelva al ejercicio 22, en el que se emplearon los datos siguientes para determinar si el precio de una impresora estaba relacionado con su velocidad para imprimir un texto (*PC World,* febrero 2003).


Printers

Nombre	Tipo	Velocidad (ppm)	Precio ($)
Minolta-QMS PagePro 1250W	Oficina	12	199
Brother HL-1850	Oficina	10	499
Lexmark E320	Oficina	12.2	299
Minolta-QMS PagePro 1250E	Oficina	10.3	299
HP Laserjet 1200	Oficina	11.7	399
Xerox Phaser 4400/N	Corporativa	17.8	1850
Brother HL-2460N	Corporativa	16.1	1000

(*continúa*)

Nombre	Tipo	Velocidad (ppm)	Precio ($)
IBM Infoprint 1120n	Corporativa	11.8	1387
Lexmark W812	Corporativa	19.8	2089
Oki Data B8300n	Corporativa	28.2	2200

¿Indican las evidencias que haya una relación significante entre velocidad de impresión y precio? Realice la prueba estadística apropiada y dé su conclusión. Use $\alpha = 0.05$.

31. En el ejercicio 20 con los datos sobre x = precio ($) y y = evaluación general de 10 televisores de plasma, de 42 pulgadas probadas por *Consumer Reports* se obtuvo la ecuación de regresión estimada $\hat{y} = 12.0169 + 0.0127x$. Con estos datos se obtuvieron SCE = 540.04 y STC = 982.40. Use la prueba F para determinar si el precio de los televisores de plasma, de 42 pulgadas y la evaluación general están relacionados. Use $\alpha = 0.05$.



14.6 Uso de la ecuación de regresión estimada para estimaciones y predicciones

Al usar el modelo de regresión lineal simple se hace una suposición acerca de la relación entre x y y. Después se usa el método de mínimos cuadrados para obtener una ecuación de regresión lineal simple estimada. Si existe una relación significante entre x y y y si el coeficiente de determinación indica que el ajuste es bueno, la ecuación de regresión estimada es útil para estimaciones y predicciones.

Estimación puntual

En el ejemplo de Armand's Pizza Parlors, la ecuación de regresión estimada $\hat{y} = 60 + 5x$ proporciona una estimación de la relación entre x el tamaño de la población de estudiantes y y las ventas trimestrales. Con la ecuación de regresión estimada se puede obtener una estimación puntual del valor medio de y correspondiente a un determinado valor de x o se puede predecir el valor de y que corresponde a un valor de x. Por ejemplo, supóngase que los gerentes de Armand's desean una estimación puntual de la media de las ventas trimestrales de todos los restaurantes que se encuentren cerca de campus de 10 000 estudiantes. Usando la ecuación de regresión estimada $\hat{y} = 60 + 5x$, con $x = 10$ (o 10 000 estudiantes) se obtiene $\hat{y} = 60 + 5(10) = 110$. Por lo tanto, una estimación puntual de la media de las ventas trimestrales de todos los restaurantes ubicados cerca de campus de 10 000 estudiantes es $110 000.

Ahora supóngase que los administradores de Armand´s desean predecir las ventas de un determinado restaurante ubicado cerca de Talbot College, una escuela de 10 000 estudiantes. En este caso lo que interesa no es la media correspondiente a todos los restaurantes que están cerca de campus de 10 000 estudiantes, sino únicamente predecir las ventas trimestrales de un determinado restaurante. En realidad, la estimación puntual de un solo valor de y es igual a la estimación puntual de la media de los valores de y. Por lo tanto, la predicción de las ventas trimestrales de este restaurante serán $\hat{y} = 60 + 5(10) = 110$ o $110 000.

Estimación por intervalo

Los intervalos de confianza y los intervalos de predicción indican la precisión de los resultados de la regresión. Los intervalos más estrechos proporcionan mayor precisión.

Las estimaciones puntuales no proporcionan información alguna acerca de la precisión de una estimación. Para eso es necesario obtener estimaciones por intervalo que son muy parecidas a las de los capítulos 8, 10 y 11. El primer tipo de estimación por intervalo, el **intervalo de confianza** es una estimación por intervalo del *valor medio de las* y que corresponden a un valor dado de x. El segundo tipo de estimación por intervalo, el **intervalo de predicción**, se usa cuando se necesita una estimación por intervalo de un *solo valor de* y para un valor dado de x. La estimación puntual del valor medio de y es igual a la estimación puntual de un solo valor de y. Pero las estimaciones por intervalo que se obtienen para estos dos casos son diferentes. En un intervalo de predicción el margen de error es mayor.

Intervalo de confianza para el valor medio de y

Con la ecuación de regresión estimada se obtiene una estimación puntual del valor medio de y que corresponde a un valor dado de x. Para obtener un intervalo de confianza se usa la notación siguiente.

x_p = valor dado de la variable independiente x

y_p = valor de la variable dependiente y que corresponde al valor dado x_p

$E(y_p)$ = valor medio o valor esperado de la variable dependiente y que corresponde al valor dado x_p

$\hat{y}_p = b_0 + b_1 x_p$ = estimación puntual de $E(y_p)$ cuando $x = x_p$

Empleando esta notación para estimar la media de las ventas de los restaurantes Armand's que se encuentran cerca de un campus de 10 000 estudiantes, se tiene que $x_p = 10$ y $E(y_p)$ denota el valor medio desconocido de las ventas de todos los restaurantes para los que $x_p = 10$. La estimación puntual de $E(y_p)$ está dada por $\hat{y}_p = 60 + 5(10) = 110$.

En general, no se puede esperar que $\hat{y}_p$ sea exactamente igual a $E(y_p)$. Para hacer una inferencia acerca de qué tan cerca está $\hat{y}_p$ de la media $E(y_p)$, es necesario estimar la varianza de $\hat{y}_p$. La fórmula para estimar la varianza de $\hat{y}_p$ para un x_p dado se denota $s^2_{\hat{y}_p}$, y es

$$s^2_{\hat{y}_p} = s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right] \tag{14.22}$$

Una estimación de la desviación estándar de $\hat{y}_p$ está dada por la raíz cuadrada de la ecuación (14.22).

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \tag{14.23}$$

En los resultados obtenidos en la sección 14.5 para el ejemplo de Armand's Pizza Parlors se tiene $s = 13.829$. Como $x_p = 10$, $\bar{x} = 14$ y $\Sigma(x_i - \bar{x})^2 = 568$, usando la ecuación (14.23) se obtiene

$$\begin{aligned} s_{\hat{y}_p} &= 13.829\sqrt{\frac{1}{10} + \frac{(10 - 14)^2}{568}} \\ &= 13.829\sqrt{0.1282} = 4.95 \end{aligned}$$

A continuación se presenta la fórmula general para obtener un intervalo de confianza.

El margen de error en esta estimación por intervalo (este intervalo de estimación) es $t_{\alpha/2}s_{\hat{y}_p}$.

INTERVALO DE CONFIANZA PARA $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2}s_{\hat{y}_p} \tag{14.24}$$

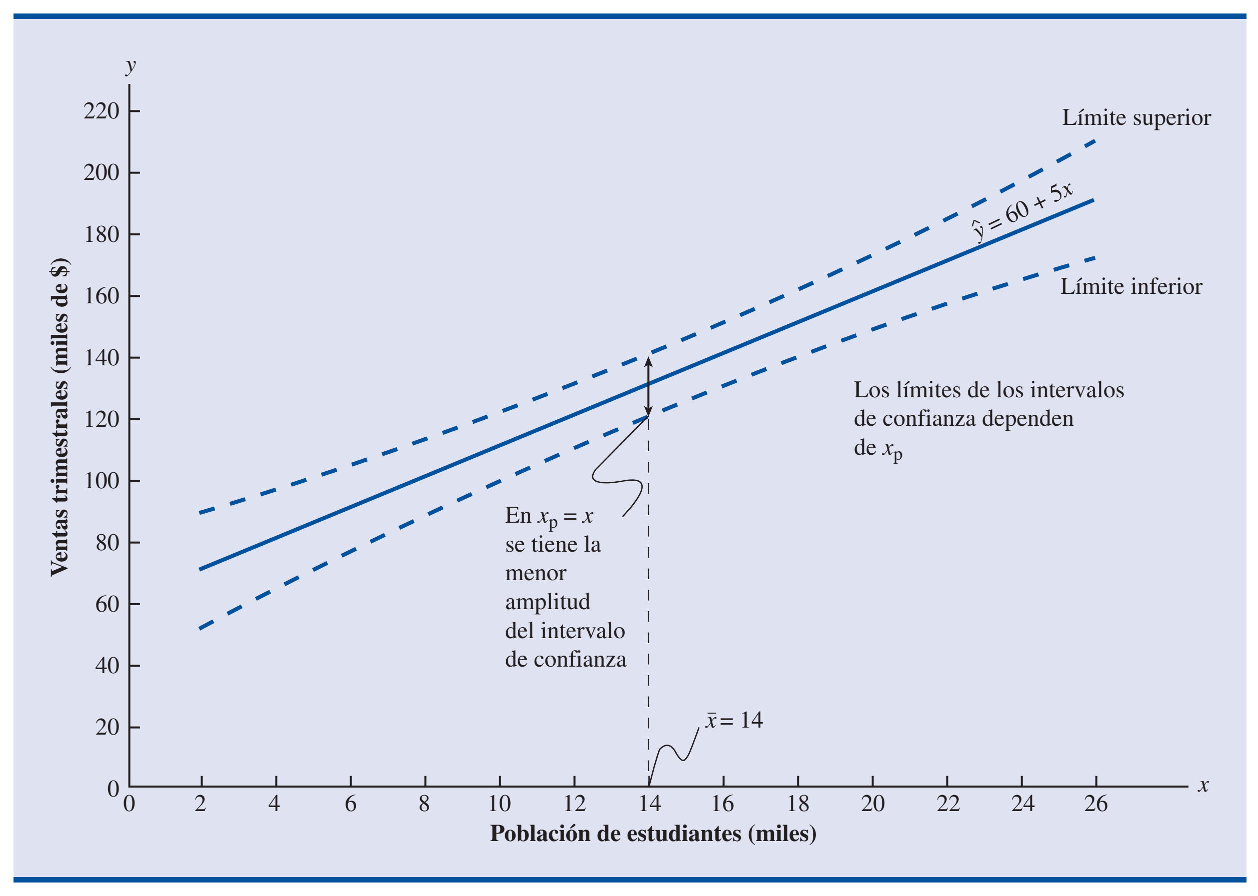
donde el coeficiente de confianza es $1 - \alpha$ y $t_{\alpha/2}$ es un valor de la distribución t con $n - 2$ grados de libertad.

Para obtener, con la fórmula (14.24), un intervalo de confianza de 95% para la media de las ventas trimestrales de los restaurantes Armand's que se encuentran cerca de campus de 10 000 estudiantes, se necesita el valor de t para $\alpha/2 = 0.025$ y $n - 2 = 10 - 2 = 8$ grados de libertad. En la tabla 2 del apéndice B, se encuentra $t_{0.025} = 2.306$. Por lo tanto, como $\hat{y}_p = 110$ y el margen de error es $t_{\alpha/2}s_{\hat{y}_p} = 2.306(4.95) = 11.415$, la estimación por intervalo de 95% de confianza es

$$110 \pm 11.415$$



FIGURA 14.8 INTERVALOS DE CONFIANZA PARA LA MEDIA DE LAS VENTAS y CORRESPONDIENTES A VALORES DADOS DEL TAMAÑO DE LA POBLACIÓN DE ESTUDIANTES x



En dólares, el intervalo de 95% de confianza para la media de las ventas trimestrales de todos los restaurantes que se encuentran cerca de un campus de 10 000 estudiantes es 110 000 $\pm$ \$11 415. Por lo tanto, si el tamaño de la población de estudiantes es 10 000, el intervalo de 95% de confianza para la media de las ventas trimestrales en los restaurantes cercanos a un campus de 10 000 estudiantes es el intervalo que va de \$98 585 a \$121 415.

Obsérvese que la desviación estándar estimada de $\hat{y}_p$, dada por la ecuación (14.23), es menor cuando $x_p = \bar{x}$ y la cantidad $x_p - \bar{x} = 0$. En este caso, la desviación estándar estimada de $\hat{y}_p$ se convierte en

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(\bar{x} - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = s\sqrt{\frac{1}{n}}$$

Esto significa que cuando $x_p = \bar{x}$ se obtiene la mejor estimación o la estimación más precisa del valor medio de y. Entre más alejada esté x_p de $\bar{x}$, mayor será $x_p - \bar{x}$. El resultado es que los intervalos de confianza para el valor medio de y son más amplios a medida que x_p se aleja de $\bar{x}$. En la figura 14.8 se muestra esto gráficamente.

Intervalo de predicción para un solo valor de *y*

Supóngase que en lugar de que lo que interese sea estimar el valor medio de las ventas de todos los restaurantes Armand's que se encuentran cerca de campus de 10 000 estudiantes, se deseen estimar las ventas de un solo restaurante que se encuentra cerca de Talbot College, una escuela

de 10 000 estudiantes. Como ya se indicó, la estimación puntual de y_p, el valor de y que corresponde a un valor dado x_p, se obtiene mediante la ecuación de regresión $\hat{y}_p = b_0 + b_1x_p$. En el caso del restaurante cerca de Talbot College, como $x_p = 10$, las ventas trimestrales pronosticadas serán $\hat{y}_p = 60 + 5(10) = 110$ o \$110 000. Obsérvese que este valor es el mismo que el obtenido como estimación puntual de la media de las ventas en los restaurantes que se encuentran cerca de campus de 10 000 estudiantes.

Para obtener un intervalo de predicción, es necesario determinar primero la varianza correspondiente al uso de $\hat{y}_p$ como estimación de un valor individual de y cuando a $x = x_p$. Esta varianza está formada por la suma de los dos componentes siguientes.

1. La varianza de los valores individuales de y respecto a la media $E(y_p)$, para la cual una estimación está dada por s^2
2. La varianza correspondiente al uso de $\hat{y}_p$ para estimar $E(y_p)$, para la cual una estimación está dada por $s^2_{\hat{y}_p}$

La fórmula para estimar la varianza de un valor individual de y_p que se denota s^2_{ind}, es

$$\begin{aligned} s^2_{\text{ind}} &= s^2 + s^2_{\hat{y}_p} \\ &= s^2 + s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right] \\ &= s^2\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right] \end{aligned} \tag{14.25}$$

Por lo tanto, una estimación de la desviación estándar de un solo valor de y_p es la dada por

$$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \tag{14.26}$$

En el ejemplo de Armand's Pizza Parlors, la desviación estándar estimada que corresponde a la predicción de las ventas de un determinado restaurante que esté cerca de un campus de 10 000 estudiantes se calcula como sigue.

$$\begin{aligned} s_{\text{ind}} &= 13.829\sqrt{1 + \frac{1}{10} + \frac{(10 - 14)^2}{568}} \\ &= 13.829\sqrt{1.1282} \\ &= 14.69 \end{aligned}$$

La fórmula general para un intervalo de predicción es como sigue

El margen de error de este intervalo de estimación es $t_{\alpha/2}s_{\text{ind}}$.

INTERVALO DE PREDICCIÓN PARA y_p

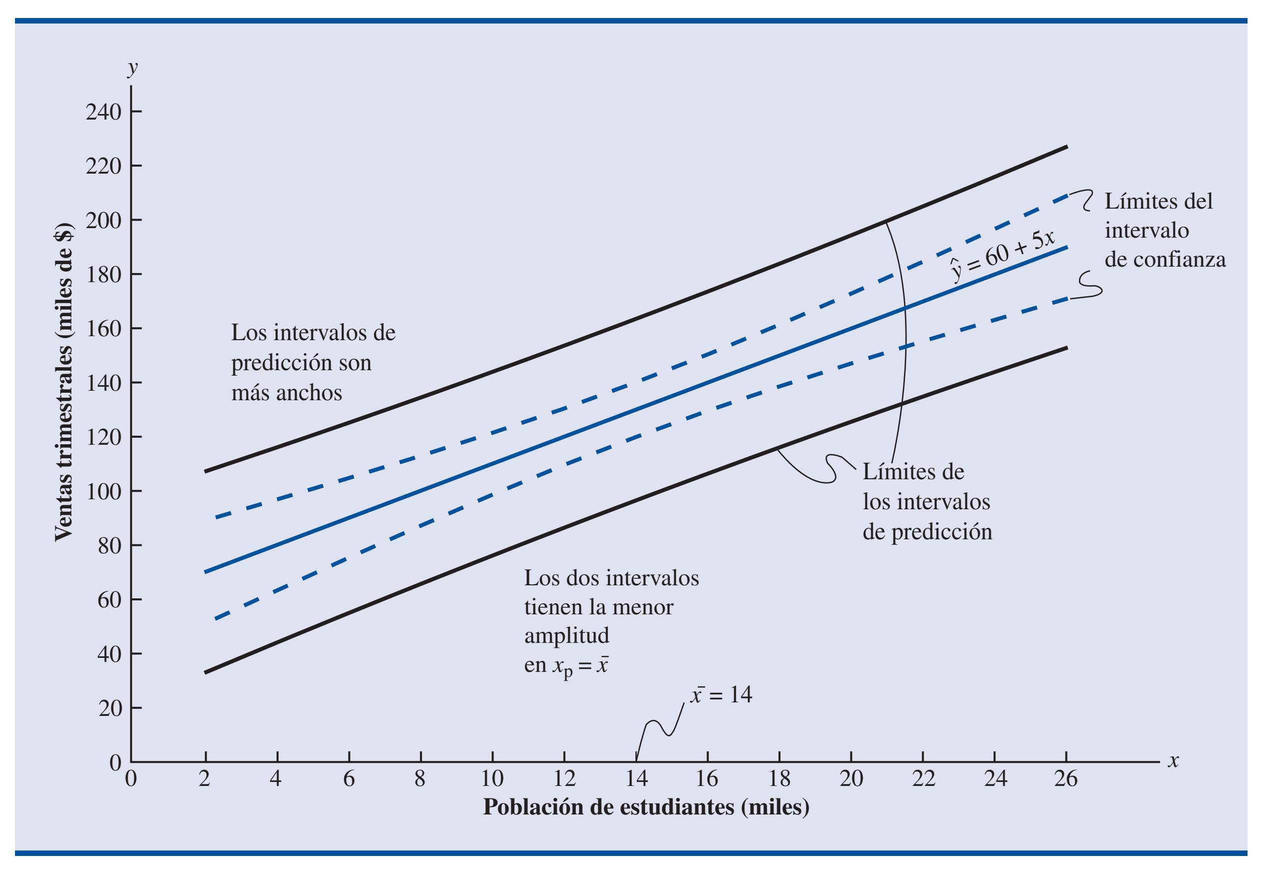
$$\hat{y}_p \pm t_{\alpha/2}s_{\text{ind}} \tag{14.27}$$

donde el coeficiente de confianza es $1 - \alpha$ y $t_{\alpha/2}$ es un valor de la distribución t para $n - 2$ grados de libertad

El intervalo de predicción de las ventas trimestrales del restaurante situado cerca de Talbot Collage se encuentra empleando $t_{0.025} = 2.306$ y $s_{\text{ind}} = 14.69$. Por lo tanto, como $\hat{y}_p = 110$ y el margen de error es $t_{\alpha/2}s_{\text{ind}} = 2.306(14.69) = 33.875$, el intervalo de predicción de 95% de confianza es

$$110 \pm 33.875$$

FIGURA 14.9 INTERVALOS DE CONFIANZA Y DE PREDICCIÓN PARA LAS VENTAS y QUE CORRESPONDEN A VALORES DADOS x DEL TAMAÑO DE LA POBLACIÓN DE ESTUDIANTES



En dólares, el intervalo de predicción es \$110 000 ± \$33 875 o el intervalo que va de \$76 125 a \$143 875. Obsérvese que el intervalo de predicción para un solo restaurante que se encuentre cerca de un campo de 10 000 estudiantes es más amplio que el intervalo de confianza para la media de las ventas de todos los restaurantes que se encuentran cerca de campus de 10 000 estudiantes. Esta diferencia refleja el hecho de que se puede estimar con más precisión la media de y que un solo valor individual de y.

En general, tanto las líneas de los límites para los intervalos de confianza como las de los límites para los intervalos de predicción tienen cierta curvatura.

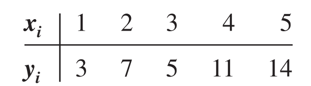
Tanto las estimaciones mediante un intervalo de confianza como las estimaciones mediante un intervalo de predicción son más precisas cuando el valor de la variable independiente es $x_p = \bar{x}$. En la figura 14.9 se muestra la forma general de los intervalos de confianza y de los intervalos de predicción que son más anchos.

Ejercicios

Métodos

Autoexamen

32. Los datos siguientes son los del ejercicio 1.



x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Use la ecuación (14.23) para estimar la desviación estándar de $\hat{y}_p$ cuando $x = 4$.
b. Use la expresión (14.24) para obtener un intervalo de confianza de 95% para el valor esperado de y cuando $x = 4$.

c. Use la ecuación (14.26) para estimar la desviación estándar de un valor de y cuando $x = 4$.
d. Use la expresión (14.27) para obtener un intervalo de predicción de 95% para y cuando $x = 4$

33. Los datos siguientes son los del ejercicio 2.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a. Estime la desviación estándar de $\hat{y}_p$ cuando $x = 8$.
b. Obtenga un intervalo de 95% de confianza para el valor esperado de y cuando $x = 8$.
c. Estime la desviación estándar de un valor individual de y cuando $x = 8$.
d. Obtenga un intervalo de predicción de 95% para y cuando $x = 8$.

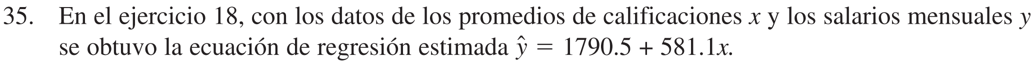
34. Los datos siguientes son los del ejercicio 3.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

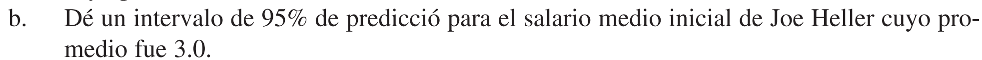
Obtenga los intervalos de confianza y de predicción del 95% para $x = 12$. Explique por qué son diferentes estos dos intervalos.

Aplicaciones

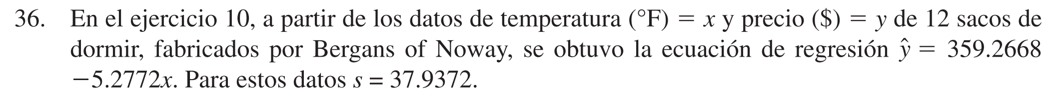



35. En el ejercicio 18, con los datos de los promedios de calificaciones x y los salarios mensuales y se obtuvo la ecuación de regresión estimada $\hat{y} = 1790.5 + 581.1x$.


a. Dé un intervalo de 95% de confianza para el salario medio inicial de todos los estudiantes cuyo promedio fue 3.0.


b. Dé un intervalo de 95% de predicció para el salario medio inicial de Joe Heller cuyo promedio fue 3.0.




36. En el ejercicio 10, a partir de los datos de temperatura (°F) = x y precio (\$) = y de 12 sacos de dormir, fabricados por Bergans of Noway, se obtuvo la ecuación de regresión $\hat{y} = 359.2668 - 5.2772x$. Para estos datos $s = 37.9372$.


a. Dé una estimación puntual del precio de un saco de dormir cuya temperatura sea 30.
b. Dé un intervalo de 95% de confianza para el precio medio de todos los sacos de dormir cuya temperatura sea 30.
c. Suponga que Bergans elabora un nuevo modelo cuya temperatura es 30. Dé un intervalo de predicción de 95% para el precio de este nuevo modelo.
d. Explique la diferencia entre sus respuestas a los incisos b) y c).

37. En el ejercicio 13 se proporcionaron datos sobre el ingreso bruto ajustado y el monto de las deducciones en las declaraciones de impuestos. Los datos se dieron en miles de dólares. Como la ecuación de regresión estimada es $\hat{y} = 4.68 + 0.16x$, el monto razonable de las deducciones, para un contribuyente cuyo ingreso bruto ajustado sea \$52 500, es \$13 080.
a. Dé un intervalo de 95% de confianza para el monto medio de las deducciones de todos los contribuyentes cuyo ingreso bruto ajustado sea \$52 500.
b. Dé un intervalo de predicción de 95% para el monto total de deducciones de un contribuyente cuyo ingreso bruto ajustado sea \$52 500.
c. Si el contribuyente del inciso b) solicita deducciones de \$20 400, ¿se justificaría que se le quiera hacer una auditoría?
d. Emplee su respuesta al inciso b) para indicar el monto de las deducciones que puede solicitar un contribuyente cuyo ingreso bruto ajustado sea \$52 500 sin que se le haga una auditoría.

38. Retome el ejercicio 21, en el que la ecuación de regresión estimada $\hat{y} = 1246.67 + 7.6x$ se obtuvo empleando los datos de volumen de producción x y costos totales y de una determinada operación de fabricación.
a. En el plan de producción de la empresa se ve que el mes próximo deberán producirse 500 unidades. Dé la estimación puntual de los costos totales.

b. Dé un intervalo de predicción de 99% para el costo total de producción de las 500 unidades, el mes próximo.
c. Si al final del mes próximo, el informe de costos de un contador indica que en ese mes los costos reales de producción fueron $6000, ¿debería preocupar a los gerentes el haber incurrido ese mes en costos totales tan altos? Analice.

39. En Estados Unidos casi todo el sistema de tranvías usa vagones eléctricos que corren sobre vías a nivel de la calle. La Administración de Tránsito Federal afirma que el tranvía es uno de los medios de transporte más seguros, ya que la tasa de accidentes es 0.99 accidentes por millón de millas-pasajero en comparación con 2.29 en los autobuses. En los datos siguientes se dan las millas de vía y la cantidad de pasajeros transportados en los días laborables, en miles, de seis sistemas de tranvías (*USA Today,* 7 de enero 2003).

Ciudad	Millas de vías	Pasajeros transportados (miles)
Cleveland	15	15
Denver	17	35
Portland	38	81
Sacramento	21	31
San Diego	47	75
San Jose	31	30
St. Louis	34	42

a. Use estos datos para obtener la ecuación de regresión estimada que podría emplearse para predecir la cantidad de pasajeros dadas las millas de vías.
b. ¿Proporciona la ecuación de regresión estimada un buen ajuste? Explique.
c. Obtenga un intervalo de 95% de confianza para la media de la cantidad de pasajeros transportados en los días laborables en los sistemas de tranvías que tienen 30 millas de vías.
d. Suponga que Charlotte está considerando la construcción de un sistema de tranvía de 30 millas de vías. Dé un intervalo de predicción de 95% para la cantidad de pasajeros transportada en un día laborable por el sistema Charlotte. ¿Cree usted que el intervalo de predicción que desarrolló pueda ser útil a los que están planeando Charlotte para anticipar la cantidad de pasajeros en un día laborable en su sistema de tranvía? Explique



Solución por computadoras

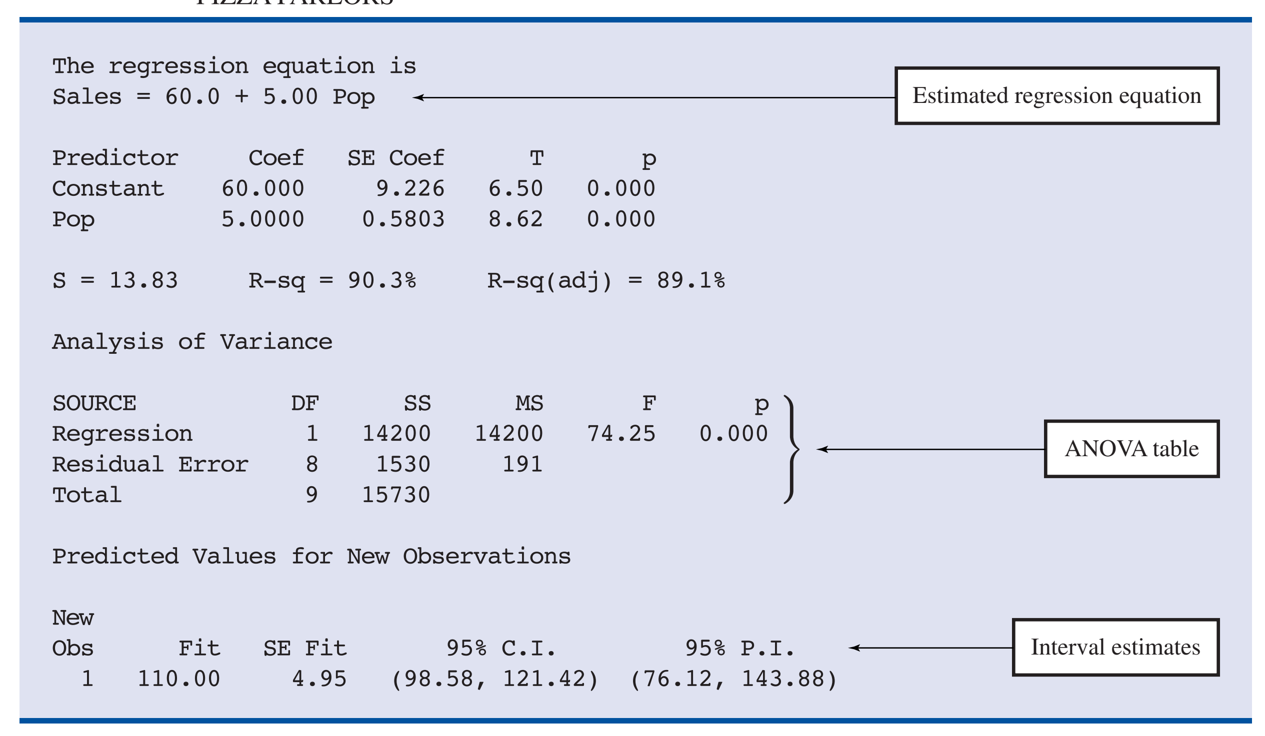
Realizar los cálculos del análisis de regresión sin la ayuda de una computadora puede costar mucho tiempo. En esta sección se verá cómo mediante el uso de paquetes de software como Minitab puede minimizarse la complicación de hacer tantos cálculos.

Los datos sobre población de estudiantes y ventas se han ingresado en la hoja de cálculo de Minitab. A la variable independiente se le ha llamado Pop y a la variable independiente se le ha llamado Ventas para facilitar la interpretación de los resultados que proporciona la computadora. Usando Minitab para el ejemplo de Armand's Pizza Parlors se obtuvieron los resultados que se muestran en la figura 14.10*. A continuación se explica cómo interpretar estos resultados.

1. Minitab da la ecuación de regresión estimada como Ventas = 60.0 + 5.00 Pop.
2. Da también una tabla en la que indica el valor de los coeficientes b_0 y b_1, la desviación estándar de cada coeficiente, el valor t obtenido al dividir cada coeficiente entre su desviación estándar y el valor-p correspondiente a la prueba t. Como el valor-p es cero (a tres cifras decimales), los resultados muestrales indican que debe rechazarse la hipótesis nula (H_0: $\beta_1 = 0$). O bien, se puede comparar 8.62 (que aparece en la columna t) con el valor crítico apropiado. Este procedimiento para la prueba t se describió en la sección 14.5.

*En el apéndice 14.3 se dan los pasos que hay que seguir con Minitab para obtener estos resultados.

FIGURA 14.10 RESULTADOS DADOS POR MINITAB PARA EL PROBLEMA DE ARMAND'S PIZZA PARLORS



```
The regression equation is
Sales = 60.0 + 5.00 Pop

Predictor      Coef    SE Coef      T       p
Constant     60.000      9.226   6.50   0.000
Pop          5.0000     0.5803   8.62   0.000

S = 13.83     R-sq = 90.3%     R-sq(adj) = 89.1%

Analysis of Variance

SOURCE            DF       SS       MS       F       p
Regression         1    14200    14200   74.25   0.000
Residual Error     8     1530      191
Total              9    15730

Predicted Values for New Observations

New
Obs      Fit    SE Fit       95% C.I.          95% P.I.
  1   110.00      4.95   (98.58, 121.42)   (76.12, 143.88)
```


3. Minitab da el error estándar de estimación, s = 13.83, así como información acerca de la bondad de ajuste. Observe que "R-sq = 90.3%" es el coeficiente de determinación expresado como porcentaje. El valor "R-sq (adj) = 89.1%" se verá en el capítulo 15.
4. La tabla ANOVA se presenta bajo el encabezado Analysis of Variance. Minitab usa el rótulo Residual Error para la fuente de variación del error. Obsérvese que DF es la abreviación de degrees of freedom (= grados de libertad) y que CMR está dado como 14 200 y ECM como 191. El cociente de estos dos valores da el valor F que es 74.25 y el correspondiente valor-p 0.000. como el valor-p es cero (a tres lugares decimales), la relación entre ventas (Sales) y población (Pop) se considera estadísticamente significante.
5. La estimación de las ventas esperadas mediante un intervalo de confianza de 95% y la estimación de las ventas de un determinado restaurante cercano a un campus de 10 000 estudiantes mediante un intervalo de estimación de 95% se dan abajo de la tabla ANOVA. El intervalo de confianza es (98.58, 121.42) y el intervalo de predicción es (76.12, 143.88) como se indicó en la sección 14.6.

Ejercicios

Aplicaciones



40. La división comercial de una empresa inmobiliaria realiza un análisis de regresión de la relación entre x, rentas brutas anuales (en miles de dólares) y y, precio de venta (en miles de dólares) de edificios de departamentos. Se obtuvieron datos sobre varias propiedades vendidas últimamente y con la computadora se obtuvieron los resultados siguientes.

```
The regression equation is
Y = 20.0 + 7.21 X

Predictor       Coef    SE Coef       T
Constant      20.000     3.2213    6.21
X              7.210     1.3626    5.29

Analysis of Variance

SOURCE            DF         SS
Regression         1    41587.3
Residual Error     7
Total              8    51984.1
```

a. ¿Cuántos edificios de departamentos había en la muestra?
b. Dé la ecuación de regresión estimada
c. ¿Cuál es el valor de s_{b_1}?
d. Use el estadístico F para probar la significancia de la relación empleando 0.05 como nivel de significancia.
e. Estime el precio de venta de un edificio de departamentos cuyas rentas anuales brutas son $50 000.

41. A continuación se presenta una parte de los resultados por computadora de un análisis de regresión en el que se relaciona $y =$ gastos de mantenimiento (dólares por mes) con x uso (horas por semana) para una marca determinada de terminal de computadora.

```
The regression equation is
Y = 6.1092 + .8951 X

Predictor       Coef     SE Coef
Constant      6.1092      0.9361
X             0.8951      0.1490

Analysis of Variance

SOURCE            DF         SS          MS
Regression         1    1575.76     1575.76
Residual Error     8     349.14       43.64
Total              9    1924.90
```

a. Dé la ecuación de regresión estimada.
b. Use una prueba t para determinar si los gastos mensuales de mantenimiento están relacionados con el uso, empleando 0.05 como nivel de significancia.
c. Utilice la ecuación de regresión estimada para predecir los gastos mensuales de mantenimiento de una terminal que se usa 25 hora por semana.

42. Un modelo de regresión que relaciona x, el número de vendedores en una sucursal, con y, las ventas anuales en esa sucursal (en miles de dólares), proporcionó el siguiente resultado de computadora empleando análisis de regresión de los datos.

```
The regression equation is
Y = 20.0 + 7.21 X

Predictor      Coef    SE Coef       T
Constant     20.000     3.2213    6.21
X             7.210     1.3626    5.29

Analysis of Variance

SOURCE            DF         SS
Regression         1    41587.3
Residual Error     7
Total              8    51984.1
```

a. Dé la ecuación de regresión estimada.
b. ¿Cuántas sucursales participaron en el estudio?
c. Calcule el estadístico F y pruebe la significancia de la relación empleando 0.05 como nivel de significancia.
d. Pronostique las ventas anuales de la sucursal Menphis. En esta sucursal hay 12 vendedores.

43. Los expertos en salud recomiendan que los corredores beban 4 onzas de agua por cada 15 minutos que corran. Aunque las botellas de plástico son una buena alternativa para la mayoría de los corredores, cuando se corre todo un día a campo traviesa se requieren sistemas de hidratación que se llevan sobre la cintura o sobre la espalda. Estos sistemas de hidratación además de permitir llevar más agua permiten llevar también alimento o ropas. Por supuesto, a medida que aumenta la capacidad de estos sistemas, aumenta también su peso y su precio. En la lista siguiente se da peso y precio de 26 de estos sistemas de hidratación (*Trail Runner Gear Guide,* 2003).



Hydration1

Modelo	**Peso (onzas)**	**Precio ($)**
Fastdraw	3	10
Fastdraw Plus	4	12
Fitness	5	12
Access	7	20
Access Plus	8	25
Solo	9	25
Serenade	9	35
Solitaire	11	35
Gemini	21	45
Shadow	15	40
SipStream	18	60
Express	9	30
Lightning	12	40
Elite	14	60
Extender	16	65
Stinger	16	65
GelFlask Belt	3	20
GelDraw	1	7
GelFlask Clip-on Holster	2	10
GelFlask Holster SS	1	10
Strider (W)	8	30

Modelo	Peso (onzas)	Precio ($)
Walkabout (W)	14	40
Solitude I.C.E.	9	35
Getaway I.C.E.	19	55
Profile I.C.E.	14	50
Traverse I.C.E.	13	60

a. Con estos datos obtenga una ecuación de regresión estimada que pueda ser empleada para predecir el precio de un sistema de hidratación en función de su peso.
b. Pruebe la significancia de la relación empleando 0.05 como nivel de significancia.
c. ¿Proporciona un buen ajuste la ecuación de regresión estimada?
d. Suponga que la ecuación de regresión estimada obtenida en el inciso a) también pueda usarse para sistemas de hidratación elaborados por otras empresas. Obtenga un intervalo de confianza de 95% para estimar el precio de todos los sistemas de hidratación que pesan 10 onzas.
e. Suponga que la ecuación de regresión estimada obtenida en el inciso a) también pueda usarse para sistemas de hidratación elaborados por otras empresas. Obtenga un intervalo de predicción de 95% para estimar el precio del sistema Back Draft elaborado por Eastern Mountain Sports; este sistema de hidratación pesa 10 onzas.

44. Cushman Wakefield, Inc. recoge datos sobre la tasa de desocupación en edificios de oficinas y las tasas de las rentas en mercados de Estados Unidos. Los datos siguientes dan la tasa de desocupación (%) y las tasas de rentas promedio (por pie cuadrado) en las zonas comerciales centrales de 18 mercados.

archivo en CD
OffRates

Mercado	Tasa de desocupación (%)	Tasa promedio ($)
Atlanta	21.9	18.54
Boston	6.0	33.70
Hartford	22.8	19.67
Baltimore	18.1	21.01
Washington	12.7	35.09
Philadelphia	14.5	19.41
Miami	20.0	25.28
Tampa	19.2	17.02
Chicago	16.0	24.04
San Francisco	6.6	31.42
Phoenix	15.9	18.74
San Jose	9.2	26.76
West Palm Beach	19.7	27.72
Detroit	20.0	18.20
Brooklyn	8.3	25.00
Downtown, NY	17.1	29.78
Midtown, NY	10.8	37.03
Midtown South, NY	11.1	28.64

a. Con estos datos trace un diagrama de dispersión; en el eje horizontal grafique la tasa de desocupación.
b. ¿Parece haber alguna relación entre las tasas de desocupación y las tasas de rentas?
c. Dé la ecuación de regresión para predecir la tasa promedio de renta en función de una tasa de desocupación dada.
d. Empleando como nivel de significancia 0.05 pruebe la significancia de esta relación.

e. ¿Proporciona, la ecuación de regresión estimada, un buen ajuste? Explique.
f. Pronostique la tasa de renta esperada en los mercados en los que la tasa de desocupación en zonas comerciales centrales es 25%.
g. La tasa de desocupación general en la zona comercial central de Ft. Lauderdale es 11.3%. Pronostique la tasa de renta esperada en Ft. Lauderdale.



Análisis residual: confirmación de las suposiciones del modelo

*El **análisis residual** es la herramienta principal para determinar si el modelo de regresión empleado es apropiado.*

Como ya se indicó, el *residual* de la observación *i* es la diferencia entre el valor observado de la variable dependiente (y_i) y el valor estimado de la variable dependiente ($\hat{y}_i$)

RESIDUAL DE LA OBSERVACIÓN *i*

$$y_i - \hat{y}_i \tag{14.28}$$

donde

y_i es el valor observado de la variable dependiente

$\hat{y}_i$ es el valor estimado de la variable dependiente

En otras palabras, el residual *i* es el error que resulta de usar la ecuación de regresión estimada para predecir el valor de la variable dependiente. En la tabla 14.7 se calculan estos residuales correspondientes a los datos del ejemplo de Armand's Pizza Parlors. En la segunda columna de la tabla se presentan los valores observados de la variable dependiente y en la tercera columna, los valores estimados de la variable dependiente obtenidos usando la ecuación de regresión estimada $\hat{y} = 60 + 5x$. Un análisis de los residuales correspondientes, que se encuentran en la cuarta columna de la tabla, ayuda a determinar si las suposiciones hechas acerca del modelo de regresión son adecuadas.

A continuación se revisan las suposiciones de regresión en el ejemplo de Armand's Pizza Parlors. Se supuso un modelo de regresión lineal simple

$$y = \beta_0 + \beta_1 x + \epsilon \tag{14.29}$$

TABLA 14.7 RESIDUALES EN EL EJEMPLO DE ARMAND'S PIZZA PARLORS

Población de estudiantes x_i	Ventas y_i	Ventas estimadas $\hat{y}_i = 60 + 5x_i$	Residuales $y_i - \hat{y}_i$
2	58	70	−12
6	105	90	15
8	88	100	−12
8	118	100	18
12	117	120	−3
16	137	140	−3
20	157	160	−3
20	169	160	9
22	149	170	−21
26	202	190	12

Este modelo indica que se supone que las ventas trimestrales (y) son función lineal del tamaño de la población de estudiantes (x), más un término del error ϵ. En la sección 14.4, para el término del error ϵ se hicieron las siguientes suposiciones

1. $E(\epsilon) = 0$.
2. La varianza de ϵ, que se denota σ^2, es la misma para todos los valores de x.
3. Los valores de ϵ son independientes.
4. El término del error ϵ tiene distribución normal.

Estas suposiciones son la base teórica para las pruebas t y F que se usan para determinar si la relación entre x y y es significativa y para las estimaciones, mediante intervalos de confianza y de predicción, presentadas en la sección 14.6. Si las suposiciones acerca del término del error ϵ son dudosas, puede ser que las pruebas de hipótesis acerca de la significancia de la relación de regresión y los resultados de la estimación por intervalo no sean correctos.

Los residuales proporcionan la mejor información acerca de ϵ; por lo tanto, el análisis de los residuales es muy importante para determinar si las suposiciones hechas acerca de ϵ son apropiadas. Gran parte del análisis residual se basa en examinar gráficas. En esta sección se estudiarán las siguientes gráficas de residuales.

1. La gráfica de residuales contra los valores de la variable independiente x
2. La gráfica de residuales contra los valores pronosticados para la variable dependiente $\hat{y}$
3. La gráfica de residuales estandarizados
4. La gráfica de probabilidad normal.

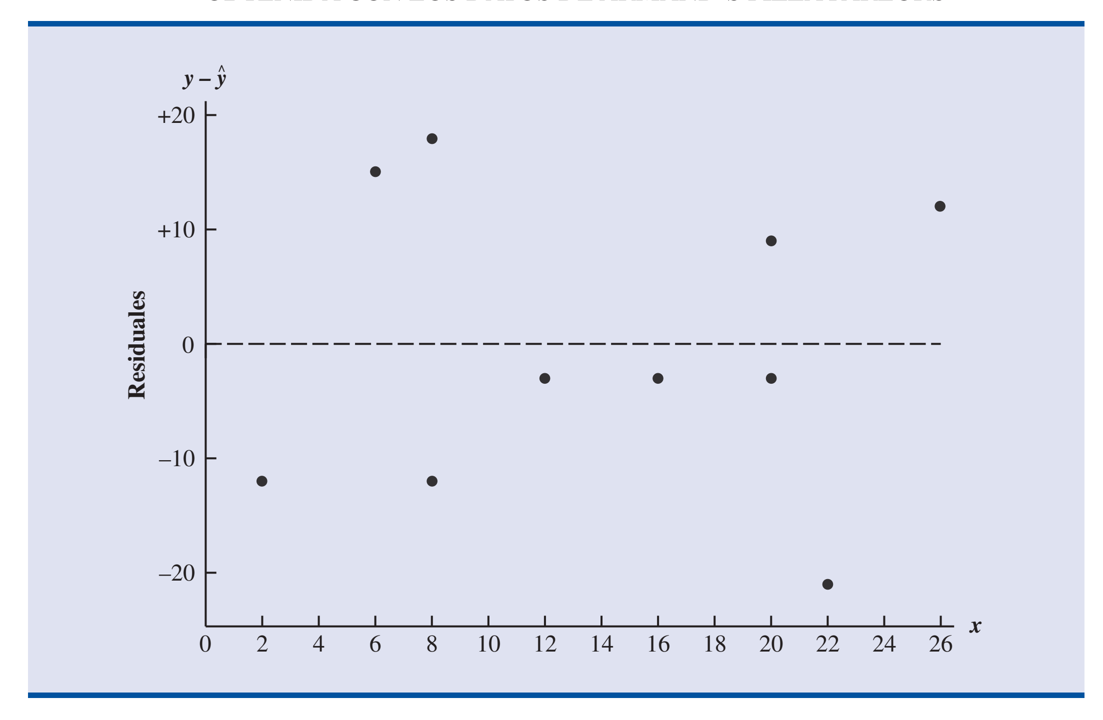
Gráfica de residuales contra x

La **gráfica de residuales** contra la variable independiente x es una gráfica en la que los valores de la variable independiente se representan en el eje horizontal y los valores de los residuales correspondientes se representan en el eje vertical. Para cada residual se grafica un punto. La primera coordenada de cada punto está dada por el valor x_i y la segunda coordenada está dada por el correspondiente valor del residual $y_i - \hat{y}_i$. En la gráfica de residuales contra x obtenida con los datos de Armand's Pizza Parlors de la tabla 14.7, las coordenadas del primer punto son (2, −12), que corresponden a $x_1 = 2$ y $y_1 - \hat{y}_1 = -12$; las coordenadas del segundo punto son (6, 15), que corresponden a $x_2 = 6$ y $y_2 - \hat{y}_2 = 15$; etc. En la figura 14.11 se muestra la gráfica de residuales obtenida.

Antes de interpretar los resultados de esta gráfica de residuales, se considerarán algunas de las formas generales que pueden tener las gráficas de residuales. En la figura 14.12 se muestran tres ejemplos. Si la suposición de que la varianza de ϵ es la misma para todos los valores de x y si el modelo de regresión empleado representa adecuadamente la relación entre las variables, el aspecto general de la gráfica de residuales será el de una banda horizontal de puntos como en la gráfica A de la figura 14.12. Pero si la varianza de ϵ no es la misma para todos los valores x —por ejemplo, si la variabilidad respecto a la línea de regresión es mayor para valores de x mayores— el aspecto de la gráfica puede ser como el de la gráfica B de la figura 14.12. En este caso, se viola la suposición de que ϵ tiene una varianza constante. En la gráfica C se muestra otra forma que puede tomar la gráfica de residuales. En este caso, se puede concluir que el modelo de regresión empleado no representa adecuadamente la relación entre las variables, y deberá considerarse un modelo de regresión curvilíneo o múltiple.

Volviendo, ahora, a la gráfica de los residuales del ejemplo de Armand's Pizza Parlors, figura 14.11. Estos residuales parecen tener una forma que se aproxima a la forma de banda horizontal de la gráfica A de la figura 14.12. Por lo tanto, se concluye que esta gráfica de residuales no muestra evidencias de que las suposiciones hechas para el modelo de regresión de Armand's puedan ser dudosas. Se concluye que el modelo de regresión lineal simple empleado para el ejemplo de Armand's, es válido.

FIGURA 14.11 GRÁFICA DE RESIDUALES CONTRA LA VARIABLE INDEPENDIENTE *x* OBTENIDA CON LOS DATOS DE ARMAND'S PIZZA PARLORS



Para la adecuada interpretación de las gráficas de residuales experiencia y criterio son muy importantes. Es raro que una gráfica de residuales tenga exactamente la forma de una de las gráficas presentadas en la figura 14.12. Sin embargo, los analistas que realizan frecuentemente estudios de regresión y gráficas de residuales se vuelven expertos en reconocer las diferencias entre las formas razonables y las que indican que se pude dudar de las suposiciones del modelo. Una gráfica de residuales proporciona una técnica para evaluar la validez de las suposiciones en un modelo de regresión.

Gráfica de residuales contra $\hat{y}$

En otra gráfica de residuales los valores pronosticados para la variable dependiente $\hat{y}$ se representan en el eje horizontal y los valores de los residuales en el eje vertical. A cada residual corresponde un punto en la gráfica. La primera coordenada de cada uno de los puntos es $\hat{y}_i$ y la segunda coordenada es el valor correspondiente del residual $i, y_i - \hat{y}_i$. Empleando los datos de Armand's, tabla 14.7, las coordenadas del primer punto son (70, −12), que corresponden a $\hat{y}_1 = 70$ y $y_1 - \hat{y}_1 = -12$; las coordenadas del segundo punto son (90, 15), etc. En la figura 14.13 se presenta esta gráfica de residuales. Obsérvese que la forma de esta gráfica de residuales es igual a la forma de la gráfica de residuales contra la variable independiente *x*. Esta no es una forma que pudiera llevar a dudar de las suposiciones del modelo. En la regresión lineal simple, tanto la gráfica de residuales contra *x* como la gráfica de residuales contra $\hat{y}$ tienen la misma forma. En el análisis de regresión múltiple, la gráfica de residuales contra $\hat{y}$ se usa más debido a que se tiene más de una variable independiente.

Residuales estandarizados

Muchas de las gráficas de residuales que se obtienen con los paquetes de software utilizan una versión estandarizada de los residuales. Como se demostró en el capítulo anterior, una variable aleatoria se estandariza sustrayéndole su media y dividiendo el resultado entre su desviación es-

FIGURA 14.12 GRÁFICAS DE LOS RESIDUALES CORRESPONDIENTES A TRES ESTUDIOS DE REGRESIONES

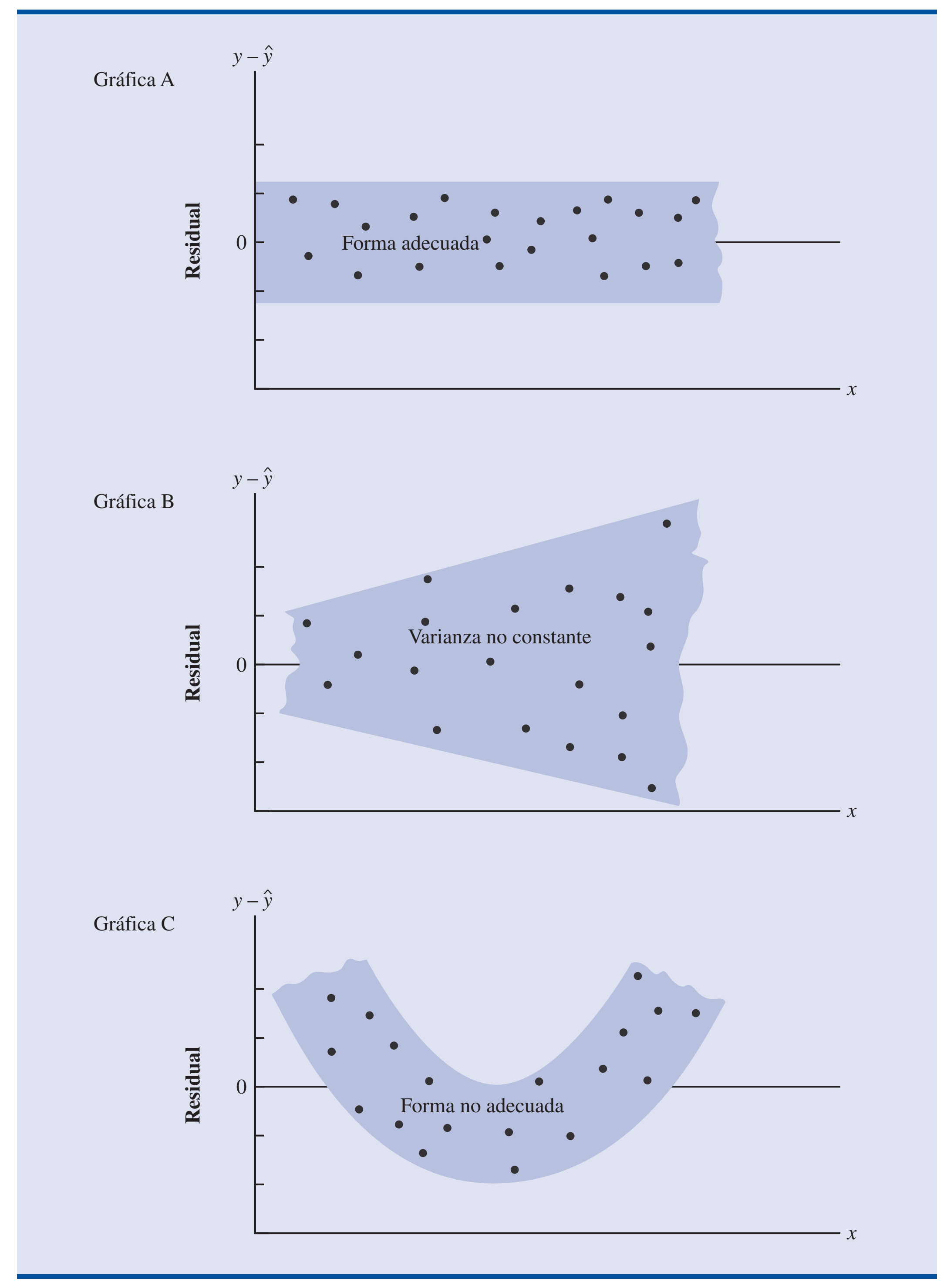
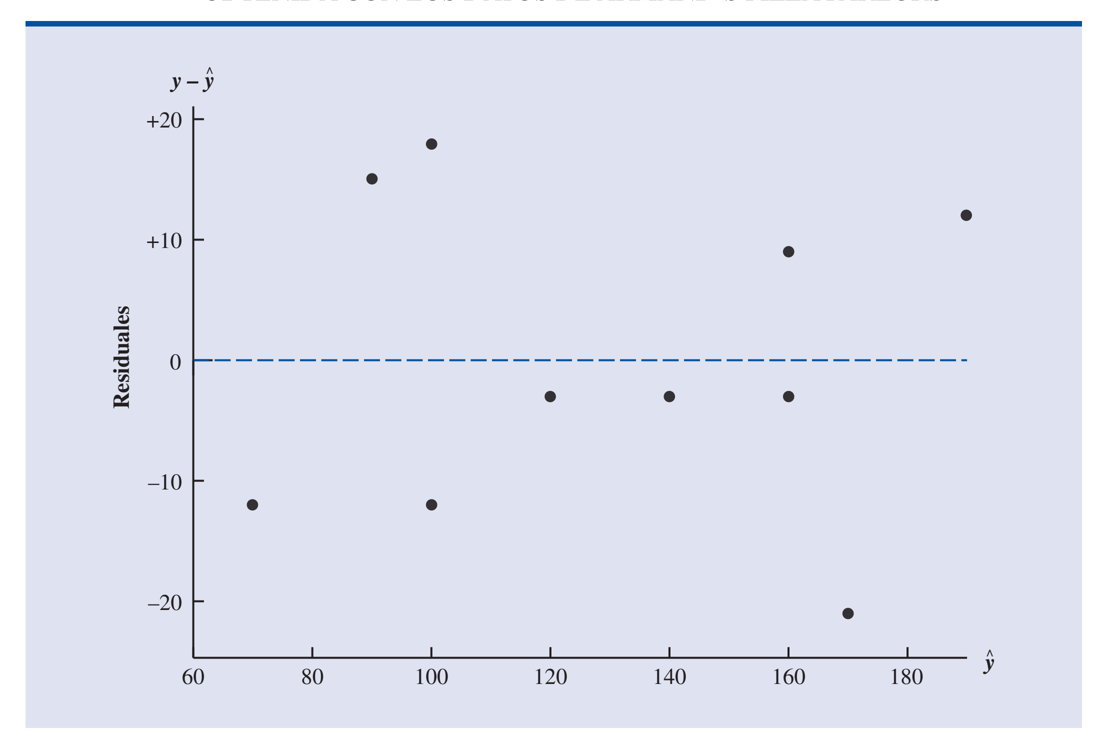


FIGURA 14.13 GRÁFICA DE RESIDUALES CONTRA EL VALOR PRONOSTICADO $\hat{y}$ OBTENIDA CON LOS DATOS DE ARMAND'S PIZZA PARLORS



tándar. Cuando se emplea el método de mínimos cuadrados, la media de los residuales es cero. Por lo tanto, para obtener el **residual estandarizado** sólo es necesario dividir cada residual entre su desviación estándar**.**

Se puede demostrar que la desviación estándar del residual i depende del error estándar de estimación s y del valor correspondiente de la variable independiente x_i.

DESVIACIÓN ESTÁNDAR DEL RESIDUAL i*

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \quad \textbf{(14.30)}$$

donde

$s_{y_i - \hat{y}_i}$ = desviación estándar del residual i

s = error estándar de estimación

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \quad \textbf{(14.31)}$$

Obsérvese que la ecuación (14.30) indica que la desviación estándar del residual i depende de x_i, debido a la presencia de h_i en la fórmula.** Una vez calculada la desviación estándar de cada uno de los residuales, se pueden calcular los residuales estandarizados dividiendo cada residual entre sus desviaciones estándar correspondientes

*En realidad, esta ecuación proporciona una estimación de la desviación estándar del residual i ya que se usa s en lugar de σ.

** A h_i se le conoce como el influencial de la observación i. El influencial se verá en la sección 14.9 cuando se consideren las observaciones influyentes.

TABLA 14.8 CÁLCULO DE LOS RESIDUALES ESTANDARIZADOS DEL EJEMPLO DE ARMAND'S PIZZA PARLORS

Restaurantes i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Residuales estandarizados
1	2	−12	144	0.2535	0.3535	11.1193	−12	−1.0792
2	6	−8	64	0.1127	0.2127	12.2709	15	1.2224
3	8	−6	36	0.0634	0.1634	12.6493	−12	−0.9487
4	8	−6	36	0.0634	0.1634	12.6493	18	1.4230
5	12	−2	4	0.0070	0.1070	13.0682	−3	−0.2296
6	16	2	4	0.0070	0.1070	13.0682	−3	−0.2296
7	20	6	36	0.0634	0.1634	12.6493	−3	−0.2372
8	20	6	36	0.0634	0.1634	12.6493	9	0.7115
9	22	8	64	0.1127	0.2127	12.2709	−21	−1.7114
10	26	12	144	0.2535	0.3535	11.1193	12	1.0792
		Total	568					

Nota: En la tabla 14.7 se calculó el valor de los residuales.

RESIDUAL ESTANDARIZADO DE LA OBSERVACIÓN i

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{14.32}$$

En la tabla 14.8 se presentan los cálculos de los residuales estandarizados utilizando el ejemplo de Armand's Pizza Parlors. Recuérdese que ya en cálculos previos se obtuvo $s = 13\ 829$. La figura 14.14 es la gráfica de los residuales estandarizados contra la variable independiente x.

Desviaciones pequeñas de la normalidad no tienen un efecto grande en las pruebas estadísticas empleadas en el análisis de regresión.

La gráfica de los residuales estandarizados permite ver si la suposición de que el término del error ϵ tiene distribución normal es correcta. Si esta suposición se satisface debe parecer que la distribución de los residuales estandarizados, proviene de una distribución de probabilidad normal estándar.* Por lo tanto, al observar la gráfica de los residuales estandarizados, se espera encontrar que aproximadamente 95% de los residuales estandarizados están entre -2 y $+2$. En la figura 14.14 se ve que en el ejemplo de Armand's todos los residuales estandarizados se encuentran entre -2 y $+2$. Por lo tanto, de acuerdo con los residuales estandarizados, esta gráfica no da razones para dudar de la suposición de que ϵ tiene una distribución normal.

Debido al trabajo que significa calcular los valores estimados de $\hat{y}$, los residuales y los residuales estandarizados, la mayoría de los paquetes de software para estadística proporcionan, de manera opcional, estos datos como parte de los resultados de la regresión.

Gráfica de probabilidad normal

Otro manera de determinar la validez de la suposición de que el término del error tiene una distribución normal es la **gráfica de probabilidad normal**. Para mostrar cómo se elabora una gráfica de probabilidad normal, se introduce el concepto de *puntos normales*.

Supóngase que, de una distribución de probabilidad normal en la que la media es cero y la desviación estándar es uno, se toman aleatoriamente 10 valores; supóngase que este proceso de muestreo se repite una y otra vez y que los 10 valores de cada muestra se ordenan de menor a mayor. Por ahora, considérese únicamente el valor menor de cada muestra. A la variable aleato-

*Como en la fórmula (14.30) se usa s en lugar de σ, la distribución de probabilidad de los residuales estandarizados no es técnicamente normal. Sin embargo, en la mayoría de los estudios de regresión, el tamaño de la muestra es suficientemente grande para que una aproximación normal sea muy buena.

FIGURA 14.14 GRÁFICA DE RESIDUALES ESTANDARIZADOS CONTRA LA VARIABLE INDEPENDIENTE x, OBTENIDA CON LOS DATOS DE ARMAND'S PIZZA PARLORS.

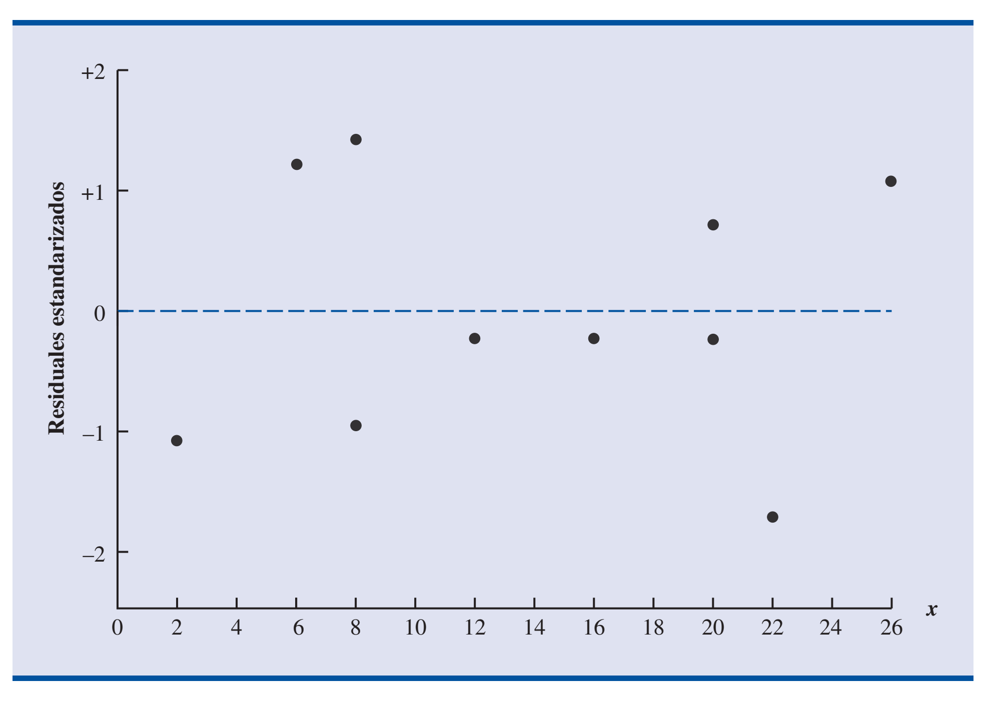


TABLA 14.9

PUNTOS NORMALES PARA $n = 10$

Estadístico de orden	Punto normal
1	−1.55
2	−1.00
3	−0.65
4	−0.37
5	−0.12
6	0.12
7	0.37
8	0.65
9	1.00
10	1.55

TABLA 14.10

PUNTOS NORMALES Y RESIDUALES ORDENADOS DE ARMAND'S PIZZA PARLORS

Puntos normales	Residuales estandarizados ordenados
−1.55	−1.7114
−1.00	−1.0792
−0.65	−0.9487
−0.37	−0.2372
−0.12	−0.2296
0.12	−0.2296
0.37	0.7115
0.65	1.0792
1.00	1.2224
1.55	1.4230

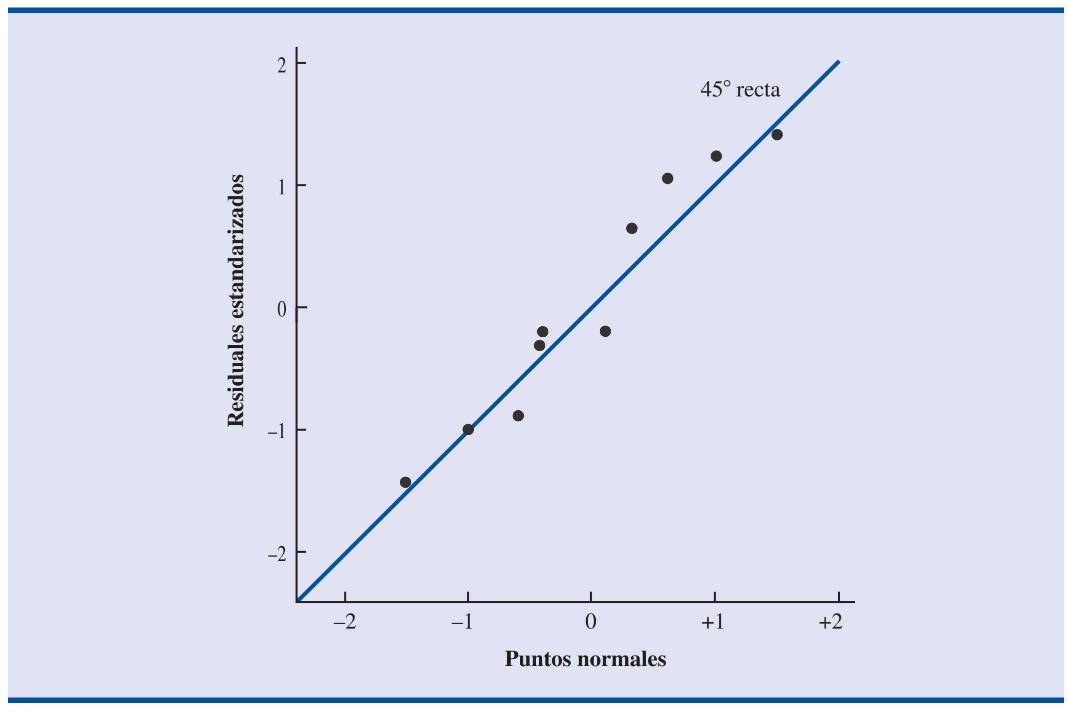
ria que representa el valor menor de estos varios muestreos se le conoce como el estadístico de primer orden.

En la ciencia de la estadística se ha demostrado que en muestras de tamaño 10 tomadas de una distribución de probabilidad normal estándar, el valor esperado del estadístico de primer orden es −1.55. A este valor esperado se le conoce como punto normal. En el caso de una muestra de tamaño $n = 10$, hay 10 estadísticos de orden y 10 puntos normales (ver tabla 14.9). En general, un conjunto de datos que conste de n observaciones tendrá n estadísticos de orden y por lo tanto n puntos normales.

A continuación se ve el uso de estos 10 puntos normales para determinar si parece ser que los residuales estandarizados de Armand's Pizza Parlors provengan de una distribución de probabilidad normal. Para empezar los 10 residuales estandarizados de la tabla 14.8 se ordenan. En la tabla 14.10 se presentan los 10 puntos normales y los residuales estandarizados normales. Si se satisface la suposición de normalidad, el menor residual estandarizado deberá tener un valor parecido al del menor punto normal, el siguiente residual estandarizado deberá tener un valor parecido al del siguiente punto normal, y así sucesivamente. En el caso de que los residuales estandarizados se encuentren distribuidos de una manera aproximadamente normal, en una gráfica en la que los puntos normales correspondan al eje horizontal y los correspondientes residuales estandarizados al eje vertical, los puntos de la gráfica estarán situados cercanos a una línea recta a 45 grados que pase por el origen. A esta gráfica es a lo que se le conoce como *gráfica de probabilidad normal*.

La figura 14.15 es la gráfica de probabilidad normal del ejemplo de Armand's Pizza Parlors. Para determinar si el patrón observado se desvía lo suficiente de la recta como para concluir que los residuales estandarizados no provienen de una distribución de probabilidad normal habrá que emplear el propio criterio. En la figura 14.15, todos los puntos se encuentran cerca de esta recta. Se concluye, por lo tanto, que la suposición de que los términos del error tienen una distribución de probabilidad normal es razonable. En general, entre más cerca de la recta a 45 grados se encuentren los puntos, más fuerte es la evidencia a favor de la suposición de normalidad. Cualquier curvatura sustancial en la gráfica de probabilidad normal es evidencia de que los residuales no provienen de una distribución de probabilidad normal. Tanto los puntos normales como la correspondiente gráfica de probabilidad normal pueden obtenerse fácilmente empleando paquetes como Minitab.

FIGURA 14.15 GRÁFICA DE PROBABILIDAD NORMAL OBTENIDA CON LOS DATOS DE ARMAND'S PIZZA PARLORS



NOTAS Y COMENTARIOS

1. Las gráficas de residuales y de probabilidad normal se usan para confirmar las suposiciones de un modelo de regresión. Si en esta revisión se encuentra que una o más de las suposiciones son dudosas, habrá que considerar otro modelo o una transformación de los datos. Cuando se violan las suposiciones, las medidas a tomar deben basarse en un criterio adecuado; las recomendaciones de una persona con experiencia en estadística pueden ser útiles.
2. El análisis de residuales es el principal método estadístico para verificar si las suposiciones del modelo de regresión son válidas. Aun cuando no se encuentre ninguna violación, esto no necesariamente implica que el modelo vaya a proporcionar buenas predicciones. Pero, si además existen otras pruebas estadísticas que favorezcan la conclusión de significancia y si el coeficiente de determinación es grande, deberá ser posible obtener buenas estimaciones y predicciones empleando la ecuación de regresión estimada.

Ejercicios

Métodos



45. Dados los datos de las dos variables *x* y *y*.

x_i	6	11	15	18	20
y_i	6	8	12	20	30

a. A partir de estos datos obtenga una ecuación de regresión estimada.
b. Calcule los residuales.
c. Trace una gráfica de residuales contra la variable independiente *x*. ¿Parecen satisfacerse las suposiciones acerca de los términos del error?

d. Calcule los residuales estandarizados.
e. Elabore una gráfica de residuales estandarizados contra $\hat{y}$. ¿Qué conclusión puede sacar de esta gráfica?

46. En un estudio de regresión se emplearon los datos siguientes.

Observación	x_i	y_i	Observación	x_i	y_i
1	2	4	6	7	6
2	3	5	7	7	9
3	4	4	8	8	5
4	5	6	9	9	11
5	7	4			

a. A partir de estos datos obtenga una ecuación de regresión estimada.
b. Trace una gráfica de residuales. ¿Parecen satisfacerse las suposiciones del término del error?

Aplicaciones



47. A continuación se presentan datos sobre los gastos en publicidad y los ingresos (en miles de dólares) del restaurante Cuatro Estaciones.

Gastos en publicidad	Ingresos
1	19
2	32
4	44
6	40
10	52
14	53
20	54

a. Sea x igual gastos en publicidad y y igual a ingresos. Utilice el método de mínimos cuadrados para obtener una línea recta que aproxime la relación entre las dos variables.
b. Empleando como nivel de significancia 0.05, pruebe si los ingresos y los gastos en publicidad están relacionados.
c. Elabore una gráfica de residuales de $y - \hat{y}$ contra $\hat{y}$. Use el resultado del inciso a) para obtener los valores de $\hat{y}$.
d. ¿Qué conclusiones se pueden sacar del análisis de residuales? ¿Se puede usar este modelo o se debe buscar uno mejor?

48. En el ejercicio 9 se obtuvo una ecuación de regresión estimada que relaciona los años de experiencia con las ventas anuales.
a. Calcule los residuales y trace una gráfica de residuales para este problema.
b. A la luz de la gráfica de residuales, ¿parecen razonables las suposiciones acerca de los términos del error?

49. American Depository Receipts (ADR) son certificados que cotizan en la bolsa de Nueva York y que representan acciones de empresas extranjeras que mantienen un depósito en un banco de su propio país. En la tabla siguiente se presenta la relación precio/ganancia (P/G) y el porcentaje de rendimiento de la inversión (ROE, por sus siglas en inglés), de 10 empresas hindúes que es probable que sean nuevos (*Bloomberg Personal Finance*, abril 2000).


ADRs

	ROE	P/G
Bharti Televentures	6.43	36.88
Gujarat Ambuja Cements	13.49	27.03
Hindalco Industries	14.04	10.83
ICICI	20.67	5.15
Mahanagar Telephone Nigam	22.74	13.35
NIIT	46.23	95.59
Pentamedia Graphics	28.90	54.85
Satyam Computer Services	54.01	189.21
Silverline Technologies	28.02	75.86
Videsh Sanchar Nigam	27.04	13.17

a. Emplee un paquete de software para obtener una ecuación de regresión estimada que relacione y = P/G y x = ROE.
b. Construya una gráfica de residuales contra la variable independiente.
c. A la luz de la gráfica de residuales, ¿parecen razonables las suposiciones acerca de los términos del error y de la forma del modelo?



14.9 Análisis de residuales: observaciones atípicas y observaciones influyentes

En la sección 14.8 se mostró cómo emplear el análisis de residuales para determinar violaciones a las suposiciones del modelo de regresión. En esta sección se ve el uso del análisis de residuales para identificar observaciones que se pueden clasificar como observaciones atípicas o como observaciones especialmente influyentes sobre la ecuación de regresión estimada. También se discuten algunas de las medidas que han de tomarse cuando se presentan tales observaciones.

Detección de observaciones atípicas

La figura 14.16 es un diagrama de dispersión de un conjunto de datos que contiene una **observación atípica**, un dato (una observación) que no sigue la tendencia del resto de los datos. Las observaciones atípicas son observaciones que son sospechosas y que requieren un análisis cuida-

FIGURA 14.16 UN CONJUNTO DE DATOS CON UNA OBSERVACIÓN ATÍPICA

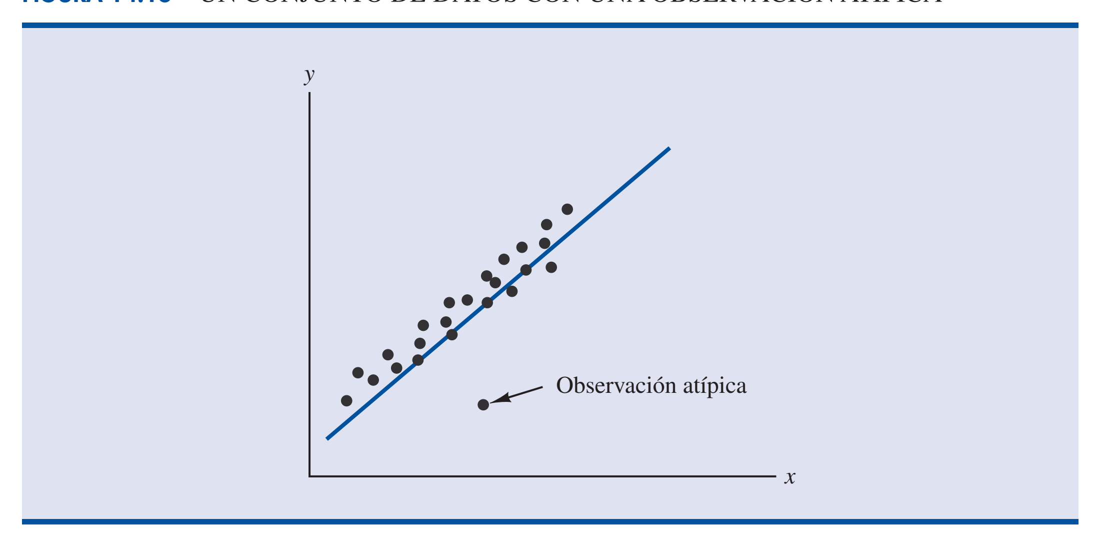


FIGURA 14.17 DIAGRAMA DE DISPERSIÓN DE UN CONJUNTO DE DATOS EN EL QUE HAY UNA OBSERVACIÓN ATÍPICA

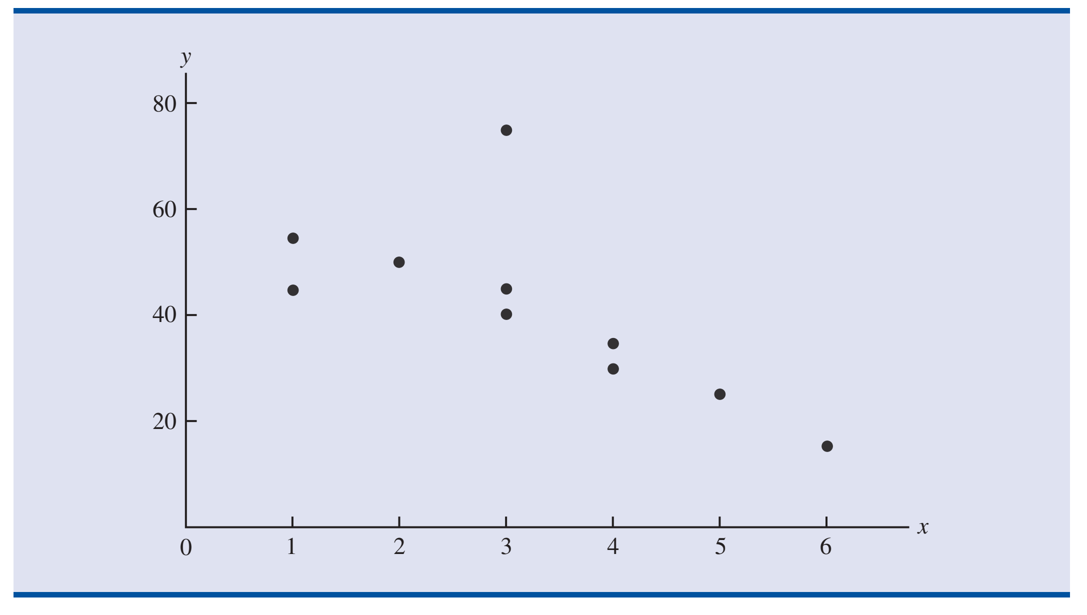


TABLA 14.11 CONJUNTO DE DATOS PARA ILUSTRAR EL EFECTO DE UNA OBSERVACIÓN ATÍPICA

x_i	y_i
1	45
1	55
2	50
3	75
3	40
3	45
4	30
4	35
5	25
6	15

doso. Puede tratarse de datos erróneos; si es así, esos datos debe ser corregidos. Puede tratarse de una violación a las suposiciones del modelo; si es así, habrá que considerar otro modelo. Por último, puede tratarse, simplemente, de valores inusuales que se presenten por casualidad. En ese caso, esos valores deberán conservarse.

Para mostrar cómo se detectan las observaciones atípicas, considérense los datos de la tabla 14.11; la figura 14.17 muestra el diagrama de dispersión de estos datos. Con excepción de la observación 4 ($x_4 = 3$, $y_4 = 75$), estos datos parecen seguir un patrón que indicar una relación lineal negativa. En efecto, dado el patrón que parece seguir el resto de los datos, se esperaría que y_4 fuera mucho más pequeño, por lo que a esta observación se le considera como un dato atípico. En el caso de la regresión lineal simple, las observaciones atípicas pueden detectarse mediante un simple examen del diagrama de dispersión.

Para detectar observaciones atípicas también se pueden usar los residuales estandarizados. Si una observación se aleja mucho del patrón del resto de los datos (por ejemplo, la observación atípica de la figura 14.16), el valor absoluto del correspondiente residual estandarizado será grande. Muchos paquetes de software identifican de manera automática las observaciones cuyos residuales tienen un valor absoluto grande. En la figura 14.18 se presentan los resultados dados por Minitab para el análisis de regresión de los datos de la tabla 14.11. En el penúltimo renglón de los resultados dados por Minitab se lee que el residual estandarizado de la observación 4 es 2.67. Minitab identifica como una observación inusual toda observación cuyo residual estandarizado sea menor a -2 o mayor a $+2$; en tales casos la observación aparece en un renglón aparte con una R al lado del residual estandarizado, como se observa en la figura 14.18. Si los errores están distribuidos normalmente, sólo 5% de los residuales estandarizados se encontrarán fuera de estos límites.

Para decidir qué hacer con una observación atípica, primero hay que verificar si es una observación correcta. Puede ser que se trate de un error al anotar los datos o al ingresarlos a la computadora. Supóngase, por ejemplo, que al verificar la observación atípica de la tabla 14.17, se encuentra que hubo un error; el valor correcto de la observación 4 era $x_4 = 3$, $y_4 = 30$. En la figura 14.19 se presenta el resultado que proporciona Minitab una vez corregido el valor de y_4. Se observa que el dato incorrecto afecta sustancialmente la bondad de ajuste. Con el dato correcto, el valor de R-sq aumenta de 49.7% a 83.8% y el valor de b_0 disminuye de 64.958 a 59.237. La pendiente de la recta cambia de -7.33 a -6.949. La identificación de los datos atípicos permite corregir errores en los datos y mejora los resultados de la regresión.

FIGURA 14.18 RESULTADOS QUE DA MINITAB PARA EL ANÁLISIS DE REGRESIÓN DEL CONJUNTO DE DATOS CON UNA OBSERVACIÓN ATÍPICA

```
The regression equation is
y = 65.0 - 7.33 x

Predictor    Coef  SE Coef      T      p
Constant   64.958    9.258   7.02  0.000
X          -7.331    2.608  -2.81  0.023

S = 12.67   R-sq = 49.7%   R-sq(adj) = 43.4%

Analysis of Variance

SOURCE          DF      SS      MS     F      p
Regression       1  1268.2  1268.2  7.90  0.023
Residual Error   8  1284.3   160.5
Total            9  2552.5

Unusual Observations
Obs     x      y    Fit  SE Fit  Residual  St Resid
  4  3.00  75.00  42.97    4.04     32.03     2.67R

R denotes an observation with a large standardized residual.
```

FIGURA 14.19 RESULTADOS QUE DA MINITAB PARA EL CONJUNTO DE DATOS CON UNA OBSERVACIÓN ATÍPICA YA CORREGIDA

```
The regression equation is
Y = 59.2 - 6.95 X

Predictor    Coef  SE Coef      T      p
Constant   59.237    3.835  15.45  0.000
X          -6.949    1.080  -6.43  0.000

S = 5.248   R-sq = 83.8%   R-sq(adj) = 81.8%

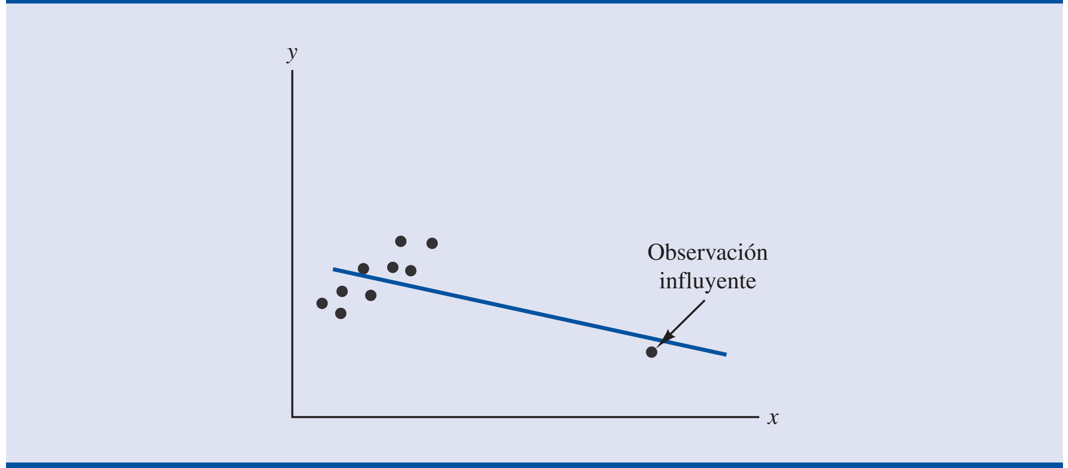
Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       1  1139.7  1139.7  41.38  0.000
Residual Error   8   220.3    27.5
Total            9  1360.0
```

Detección de observaciones influyentes

Algunas veces una o más de las observaciones tienen una influencia fuerte sobre los resultados que se obtienen. En la figura 14.20 se muestra un ejemplo de una **observación influyente** en una regresión lineal simple. La recta de regresión estimada tiene pendiente negativa, pero si la observación influyente se elimina del conjunto de datos, la pendiente de la recta de regresión estimada cambia de negativa a positiva y la intersección con el eje *y* es menor. Es claro que esta sola


FIGURA 14.20 CONJUNTO DE DATOS CON UNA OBSERVACIÓN INFLUYENTE



observación tiene mucha más influencia sobre la recta de regresión estimada que cualquiera otra observación; el efecto que tiene la eliminación de cualquiera de las otras observaciones sobre la ecuación de regresión estimada es muy pequeño.

Cuando sólo se tiene una variable independiente, las observaciones influyentes pueden identificarse mediante un diagrama de dispersión. Una observación influyente puede ser una observación atípica (una observación cuyo valor de y se desvía sustancialmente de la tendencia general), puede ser un valor de x muy alejado de la media (por ejemplo, ver la figura 14.20) o puede tratarse de la combinación de estas dos cosas (un valor de y algo fuera de la tendencia y un valor de x un poco extremo).

Las observaciones influyentes deben examinarse cuidadosamente dado el gran efecto que tienen sobre la ecuación de regresión estimada. Lo primero que hay que hacer es verificar que no se haya cometido algún error al recolectar los datos. Si se cometió algún error, se corrige y se obtiene una nueva ecuación de regresión estimada. Si la observación es correcta, puede uno considerarse afortunado de tenerla. Tal dato, cuando es correcto, contribuye a una mejor comprensión del modelo adecuado y conduce a una mejor ecuación de regresión estimada. En la figura 14.20, la presencia de la observación influyente, si es correcta, llevará a tratar de obtener datos con valores x intermedios, que permitan comprender mejor la relación entre x y y.

Las observaciones en las que la variable independiente toma valores extremos se denominan **datos (puntos, observaciones) de gran influencia**. La observación influyente de la figura 14.20 es un punto de gran influencia. La influencia de una observación depende de qué tan lejos está el valor de la variable independiente de su media. En el caso de una sola variable independiente, la influencia (*leverage*) de la observación i, que se denota h_i se calcula mediante la ecuación (14.33).

INFLUENCIA DE LA OBSERVACIÓN i

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \tag{14.33}$$

De acuerdo con esta fórmula, es claro que entre más alejada se encuentre x_i de su media $\bar{x}$ mayor será la influencia (leverage) de la observación i.

Muchos de los paquetes para estadística identifican automáticamente, como parte de los resultados estándar de regresión, los puntos de gran influencia. Para ilustrar cómo identifica Minitab los puntos de gran influencia, se considerará el conjunto de datos de la tabla 14.12.

FIGURA 14.21 DIAGRAMA DE DISPERSIÓN DEL CONJUNTO DE DATOS CON UN DATO DE GRAN INFLUENCIA

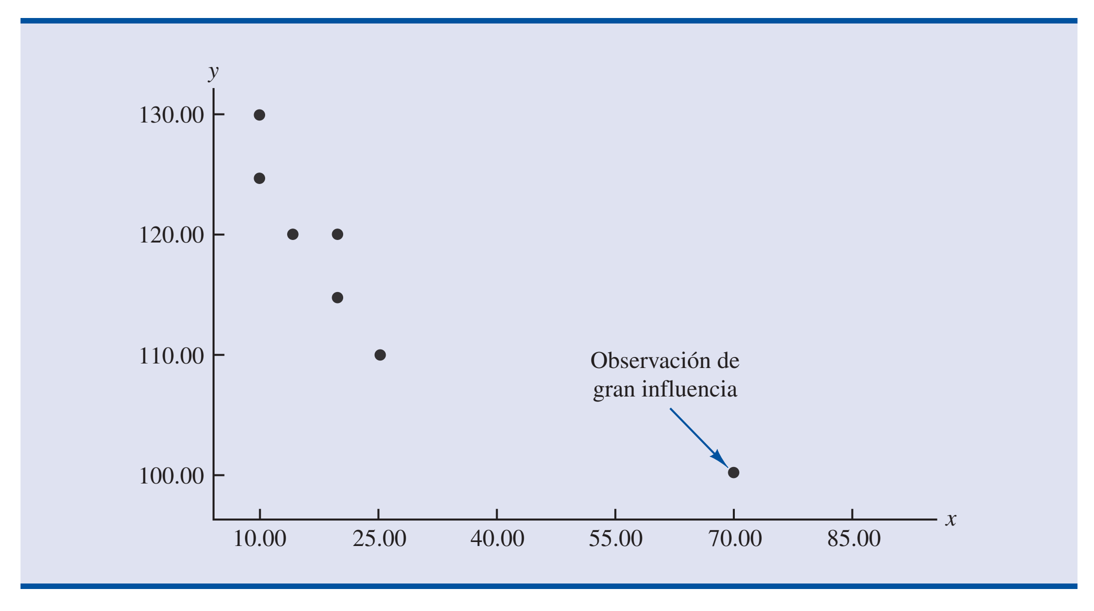


TABLA 14.12
CONJUNTO DE DATOS CON UNA OBSERVACIÓN DE GRAN INFLUENCIA

x_i	y_i
10	125
10	130
15	120
20	115
20	120
25	110
70	100

Los paquetes de software son esenciales para hacer los cálculos que permiten determinar las observaciones influyentes. Aquí se discute la regla de selección que emplea Minitab.

Observando la figura 14.21, que es el diagrama de dispersión del conjunto de datos presentado en la tabla 14.12, se ve que la observación 7 ($x = 70$, $y = 100$) es una observación en la que el valor de x es un valor extremo. Por lo tanto, es de esperarse que sea identificado como un punto de gran influencia. La influencia de esta observación se calcula usando la ecuación (14.33).

$$h_7 = \frac{1}{n} + \frac{(x_7 - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} = \frac{1}{7} + \frac{(70 - 24.286)^2}{2621.43} = 0.94$$

En el caso de la regresión lineal simple, Minitab identifica como observaciones de gran influencia las observaciones para las que $h_i > 6/n$ o $h_i = 0.99$, lo que sea menor. En el conjunto de datos de la tabla 14.12, $6/n = 6/7 = 0.86$. Como $h_7 = 0.94 > 0.86$. Minitab identificará la observación 7 como una observación cuyo valor x tiene una gran influencia. En la figura 14.22 se presenta el resultado que da Minitab del análisis de regresión de este conjunto de datos. A la observación 7 ($x = 70$, $y = 100$) la identifica como una observación de gran influencia; esta observación la presenta en un renglón aparte en la parte inferior de los resultados con una X en el margen derecho.

Las observaciones influyentes debidas a la interacción de una observación de gran influencia y de residuales grandes, suelen ser difíciles de detectar. Existen procedimientos de diagnóstico que para determinar si una observación es influyente toman en cuenta ambas cosas. En el capítulo 15 se estudiará uno de estos procedimientos, el estadístico D de Cook.

NOTAS Y COMENTARIOS

Una vez identificada una observación como potencialmente influyente, debido a que tiene un residual grande o por ser de gran influencia, su impacto sobre la ecuación de regresión estimada debe ser evaluado. En textos más avanzados se presentan los métodos de diagnóstico apropiados. Pero, cuando no se está familiarizado con el material más avanzado, una manera sencilla de hacer este diagnóstico es realizar el análisis de regresión con y sin esa observación. Este método permite apreciar la influencia que tiene la observación potencialmente influyente sobre el resultado.

FIGURA 14.22 RESULTADO DE MINITAB EMPLEANDO EL CONJUNTO DE DATOS CON UNA OBSERVACIÓN DE GRAN INFLUENCIA

```
The regression equation is
y = 127 - 0.425 x

Predictor       Coef   SE Coef      T      p
Constant     127.466     2.961  43.04  0.000
X           -0.42507   0.09537  -4.46  0.007

S = 4.883   R-sq = 79.9%   R-sq(adj) = 75.9%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       1  473.65  473.65  19.87  0.007
Residual Error   5  119.21   23.84
Total            6  592.86

Unusual Observations
Obs     x        y    Fit  SE Fit  Residual  St Resid
  7  70.0   100.00  97.71    4.73      2.29      1.91 X

X denotes an observation whose X value gives it large influence.
```

Ejercicios

Métodos



50. Considérense los datos siguientes para las variables *x* y *y*.

x_i	135	110	130	145	175	160	120
y_i	145	100	120	120	130	130	110

a. Calcule los residuales estandarizados de estos datos. ¿Hay entre los datos alguna observación atípica? Explique.
b. Haga una gráfica de residuales estandarizados contra $\hat{y}$. ¿Se observa en esta gráfica la presencia de alguna observación atípica?
c. Con estos datos elabore un diagrama de dispersión. ¿Se observa en el diagrama de dispersión la presencia de alguna observación atípica? En general, ¿qué consecuencias tienen, para la regresión lineal simple, estos hallazgos?

51. Considérense los datos siguientes para las variables *x* y *y*.

x_i	4	5	7	8	10	12	12	22
y_i	12	14	16	15	18	20	24	19

a. Calcule los residuales estandarizados de estos datos. ¿Hay entre los datos alguna observación atípica? Explique.
b. Calcule las observaciones de influencia que haya en estos datos. Entre estos datos, ¿parece haber alguna observación influyente? Explique.
c. Con estos datos elabore un diagrama de dispersión. ¿Se observa en el diagrama de dispersión la presencia de alguna observación atípica? Explique.

Aplicaciones

52. Los datos siguientes muestran los gastos (en millones de $) y los envíos en bbls. (millones) de 10 importantes marcas de cerveza.



Marca	Gastos medios (millones de $)	Envío
Budweiser	120.0	36.3
Bud Light	68.7	20.7
Miller Lite	100.1	15.9
Coors Light	76.6	13.2
Busch	8.7	8.1
Natural Light	0.1	7.1
Miller Genuine Draft	21.5	5.6
Miller High Life	1.4	4.4
Busch Light	5.3	4.3
Milwaukee's Best	1.7	4.3

a. Con estos datos obtenga una ecuación de regresión estimada.
b. Emplee el análisis residual para hallar observaciones atípicas u observaciones influyentes. Resuma sus hallazgos y conclusiones.

53. Los especialistas en salud recomiendan que las personas que corren tomen unos 200 ml de agua cada 15 minutos mientras están corriendo. Las personas que corren de tres a ocho horas, requieren sistemas de hidratación que se llevan sobre la cintura o sobre la espalda. En los datos a continuación se da el volumen (en onzas fluidas, 1 oz. flu = 30 ml aprox.) y el precio de 26 sistemas de hidratación que se llevan sobre la cintura o sobre la espalda (*Trail Runner Gear Guide,* 2003).



Modelo	Volumen (oz fl)	Precio ($)
Fastdraw	20	10
Fastdraw Plus	20	12
Fitness	20	12
Access	20	20
Access Plus	24	25
Solo	20	25
Serenade	20	35
Solitaire	20	35
Gemini	40	45
Shadow	64	40
SipStream	96	60
Express	20	30
Lightning	28	40
Elite	40	60
Extender	40	65
Stinger	32	65
GelFlask Belt	4	20
GelDraw	4	7
GelFlask Clip-on Holster	4	10
GelFlask Holster SS	4	10
Strider (W)	20	30
Walkabout (W)	230	40
Solitude I.C.E.	20	35
Getaway I.C.E.	40	55
Profile I.C.E.	64	50
Traverse I.C.E.	64	60

a. Obtenga la ecuación de regresión estimada que sirva para predecir el precio de un sistema de hidratación, dado su volumen.
b. Use el análisis residual para determinar si hay observaciones atípicas u observaciones influyentes. Resuma sus hallazgos y conclusiones.

54. En la tabla siguiente se presenta la capitalización de mercado y los salarios del presidente del consejo de administración (CEO, por sus siglas en inglés) de 20 empresas (*The Wall Street Journal*, 24 de febrero de 2000 y 6 de abril de 2000).

archivo en CD
CEO

	Capitalización de mercado (millones de $)	Salario del CEO (miles de $)
Anheuser-Busch	32 977.4	1130
AT&T	162 365.1	1400
Charles Schwab	31 363.8	800
Chevron	56 849.0	1350
DuPont	68 848.0	1000
General Electric	507 216.8	3325
Gillette	44 180.1	978
IBM	194 455.9	2000
Johnson & Johnson	143 131.0	1365
Kimberly-Clark	35 377.5	950
Merrill Lynch	31 062.1	700
Motorola	92 923.7	1275
Philip Morris	54 421.2	1625
Procter & Gamble	144 152.9	1318.3
Qualcomm	116 840.8	773
Schering-Plough	62 259.4	1200
Sun Microsystems	120 966.5	116
Texaco	30 040.7	950
USWest	36 450.8	897
Walt Disney	61 288.1	750

a. Obtenga la ecuación de regresión estimada para predecir el salario del CEO dada la capitalización de mercado.
b. Use el análisis de residuales para determinar si hay observaciones atípicas u observaciones influyentes. Resuma sus hallazgos y conclusiones.

Resumen

En este capítulo se mostró el uso del análisis de regresión para determinar cómo es la relación entre una variable dependiente y y una variable independiente x. En la regresión lineal simple, el modelo de regresión es $y = \beta_0 + \beta_1 x + \epsilon$. La ecuación de regresión lineal simple $E(y) = \beta_0 + \beta_1 x$ describe la relación de la media o valor esperado de y con x. Para obtener la ecuación de regresión estimada $\hat{y} = b_0 + b_1 x$ se emplearon datos muestrales y el método de mínimos cuadrados. En efecto, b_0 y b_1 son estadísticos muestrales que se usan para estimar los parámetros desconocidos del modelo, β_0 y β_1.

El coeficiente de determinación se presentó como una medida de la bondad de ajuste de la ecuación de regresión estimada; el coeficiente de determinación se puede interpretar como la proporción de la variación en la variable dependiente que puede ser explicada por la ecuación de regresión estimada. Se volvió a ver la correlación como una medida descriptiva de la intensidad de la relación lineal entre las dos variables.

Se discutieron las suposiciones acerca del modelo de regresión y del correspondiente término del error, y se presentaron las pruebas t y F, basadas en esas suposiciones, como un medio para determinar si la relación entre las dos variables es estadísticamente significativa. Se mostró

cómo usar la ecuación de regresión estimada para obtener estimaciones por medio de intervalos de confianza para el valor medio de *y* y estimaciones por medio de intervalos de predicción para valores individuales de *y*.

El capítulo concluyó con una sección sobre soluciones por computadora de los problemas de regresión y dos secciones sobre el uso del análisis residual para verificar las suposiciones del modelo e identificar las observaciones atípicas e influyentes.

Glosario

Variable dependiente La variable que se predice o explicada. Se denota *y*.
Variable independiente Variable que predice o explica. Se denota *x*.
Regresión lineal simple Análisis de regresión en el que participan una variable independiente y una variable dependiente, y en el que la relación entre estas variables se aproxima mediante una línea recta.
Modelo de regresión Ecuación que describe cómo están relacionadas *y* y *x*, más un término del error; en la regresión lineal simple, el modelo de regresión es $y = \beta_0 + \beta_1 x + \epsilon$.
Ecuación de regresión Ecuación que describe cómo está relacionada la media o valor esperado de la variable dependiente con la variable independiente; en la regresión lineal simple, $E(y) = \beta_0 + \beta_1 x$.
Ecuación de regresión estimada Estimación de la ecuación de regresión obtenida a partir de datos muestrales, empleando el método de mínimos cuadrados. En la regresión lineal simple, la ecuación de regresión estimada es $\hat{y} = b_0 + b_1 x$.
Método de mínimos cuadrados Procedimiento empleado para obtener la ecuación de regresión estimada. El objetivo es minimizar $\Sigma(y_i - \hat{y}_i)^2$.
Diagrama de dispersión Gráfica de datos bivariados en la que la variable independiente va en el eje horizontal y la variable dependiente va en el eje vertical.
Coeficiente de determinación Medida de la bondad de ajuste de la ecuación de regresión estimada. Se puede interpretar como la proporción de la variabilidad de la variable dependiente *y* que es explicada por la ecuación de regresión estimada.
Residual *i* Diferencia que existe entre el valor observado de la variable dependiente y el valor pronosticado empleando la ecuación de regresión estimada; para la observación *i*, el residual *i* es $y_i - \hat{y}_i$.
Coeficiente de correlación Medida de la intensidad de la relación lineal entre dos variables (ya visto en el capítulo 3).
Error cuadrado medio Estimación insesgada de la varianza del término del error σ^2. Se denota ECM o s^2.
Error estándar de estimación Raíz cuadrada del error cuadrado medio, se denota *s*. Es una estimación de σ, la desviación estándar del error.
Tabla ANOVA En el análisis de varianza, tabla que se usa para resumir los cálculos necesarios en la prueba *F* de significancia.
Intervalo de confianza Estimación por intervalo del valor medio de *y* para un valor dado de *x*.
Intervalo de predicción Estimación por intervalo de un solo valor de *y* para un valor dado de *x*.
Análisis residual Análisis de los residuales que se usa para determinar si parecen ser válidas las suposiciones hechas acerca del modelo de regresión. El análisis de residuales también se usa para identificar observaciones atípicas y observaciones influyentes.
Gráfica de residuales Representación gráfica de los residuales, se usa para determinar si parecen ser válidas las suposiciones hechas acerca del modelo de regresión.
Residual estandarizado Valor obtenido al dividir un residual entre su desviación estándar.

Gráfica de probabilidad normal Gráfica en la que los residuales estandarizados se grafican contra los puntos normales. Esta gráfica ayuda a determinar si parece ser válida la suposición de que los términos del error tienen una distribución de probabilidad normal.
Observación atípica Dato u observación que no sigue la tendencia del resto de los datos.
Observación influyente Observación en la que la variable independiente tiene un valor extremo.
Puntos de gran influencia Observaciones en las que la variable independiente tiene valores extremos.

Fórmulas clave

Modelo de regresión lineal simple

$$y = \beta_0 + \beta_1 x + \epsilon \tag{14.1}$$

Ecuación de regresión lineal simple

$$E(y) = \beta_0 + \beta_1 x \tag{14.2}$$

Ecuación de regresión lineal simple estimada

$$\hat{y} = b_0 + b_1 x \tag{14.3}$$

Criterio de mínimos cuadrados

$$\text{mín } \Sigma(y_i - \hat{y}_i)^2 \tag{14.5}$$

Intersección con el eje *y* y pendiente de la ecuación de regresión lineal simple

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \tag{14.6}$$

$$b_0 = \bar{y} - b_1\bar{x} \tag{14.7}$$

Suma de cuadrados debidos al error

$$\text{SCE} = \Sigma(y_i - \hat{y}_i)^2 \tag{14.8}$$

Suma de cuadrados total

$$\text{STC} = \Sigma(y_i - \bar{y})^2 \tag{14.9}$$

Suma de cuadrados debida a la regresión

$$\text{SCR} = \Sigma(\hat{y}_i - \bar{y})^2 \tag{14.10}$$

Relación entre STC, SCR y SCE

$$\text{STC} = \text{SCR} + \text{SCE} \tag{14.11}$$

Coeficiente de determinación

$$r^2 = \frac{\text{SCR}}{\text{STC}} \tag{14.12}$$

Coeficiente de correlación muestral

$$\begin{aligned} r_{xy} &= (\text{signo de } b_1)\sqrt{\text{Coeficiente de determinación}} \\ &= (\text{signo de } b_1)\sqrt{r^2} \end{aligned} \tag{14.13}$$

Error cuadrado medio (estimación de σ^2)

$$s^2 = \text{ECM} = \frac{\text{SCE}}{n - 2} \tag{14.15}$$

Error estándar de estimación

$$s = \sqrt{\text{CME}} = \sqrt{\frac{\text{SCE}}{n - 2}} \tag{14.16}$$

Desviación estándar de b_1

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} \tag{14.17}$$

Desviación estándar estimada de b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \tag{14.18}$$

Estadístico de prueba t

$$t = \frac{b_1}{s_{b_1}} \tag{14.19}$$

Regresión cuadrática media

$$\text{CMR} = \frac{\text{SCR}}{\text{Número de variables independientes}} \tag{14.20}$$

Estadístico de prueba F

$$F = \frac{\text{CMR}}{\text{CME}} \tag{14.21}$$

Desviación estándar estimada de $\hat{y}_p$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \tag{14.23}$$

Intervalo de confianza para $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \tag{14.24}$$

Desviación estándar estimada para un solo valor

$$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \tag{14.26}$$

Intervalo de predicción para y_p

$$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}} \tag{14.27}$$

Residual de la observación i

$$y_i - \hat{y}_i \tag{14.28}$$

Desviación estándar del residual *i*

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \tag{14.30}$$

Residual estandarizado de la observación *i*

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{14.32}$$

Influencia de la observación

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \tag{14.33}$$

Ejercicios complementarios

55. Si el valor de r^2 es elevado, ¿implica eso que entre las dos variables hay una relación de causa y efecto?
56. Explique con sus propias palabras la diferencia entre estimación por intervalo del valor medio de las y para un valor dado de x y estimación por intervalo de un valor de y para una x dada.
57. ¿Qué objeto tiene probar si $\beta_1 = 0$? Si se rechaza que $\beta_1 = 0$, ¿significa eso un buen ajuste?
58. En la tabla siguiente se da el número de acciones vendidas (en millones) y el precio esperado (el promedio del precio mínimo y del precio máximo) de 10 acciones de oferta pública inicial.

archivo en CD

IPO

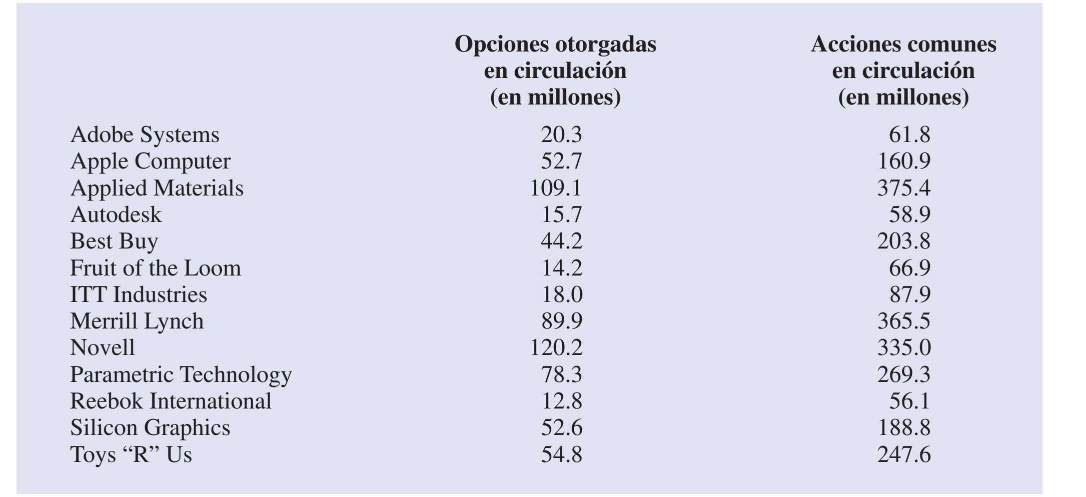
Empresa	Acciones vendidas	Precio esperado ($)
American Physician	5.0	15
Apex Silver Mines	9.0	14
Dan River	6.7	15
Franchise Mortgage	8.75	17
Gene Logic	3.0	11
International Home Foods	13.6	19
PRT Group	4.6	13
Rayovac	6.7	14
RealNetworks	3.0	10
Software AG Systems	7.7	13

a. Obtenga la ecuación de regresión estimada en la que la cantidad de acciones vendidas sea la variable independiente y el precio la variable dependiente.
b. Empleando 0.05 como nivel de significancia, ¿existe una relación significativa entre las dos variables?
c. ¿Proporciona la ecuación de regresión estimada un buen ajuste? Explique.
d. Empleando la ecuación de regresión estimada, estime el precio esperado en una empresa que considera una oferta pública inicial de 6 millones de acciones.

59. Los programas de recompra de acciones corporativas, suelen promoverse como un beneficio para los accionistas. Pero Robert Gabele, director de investigación interna de First Call/Thomson Financial, hizo notar que muchos de estos programas se realizan únicamente con el objetivo de obtener acciones que se emplean como opciones como incentivo para los altos directivos de la empresa. En todas las empresas, las opciones de acciones existentes en 1998 representaban el 6.2 por ciento de todas las acciones comunes en circulación. En los datos siguientes se da la cantidad de opciones otorgadas y la cantidad de acciones en circulación de 13 empresas (*Bloomberg Personal Finance,* enero/febrero, 2000)



Options



	Opciones otorgadas en circulación (en millones)	Acciones comunes en circulación (en millones)
Adobe Systems	20.3	61.8
Apple Computer	52.7	160.9
Applied Materials	109.1	375.4
Autodesk	15.7	58.9
Best Buy	44.2	203.8
Fruit of the Loom	14.2	66.9
ITT Industries	18.0	87.9
Merrill Lynch	89.9	365.5
Novell	120.2	335.0
Parametric Technology	78.3	269.3
Reebok International	12.8	56.1
Silicon Graphics	52.6	188.8
Toys “R” Us	54.8	247.6

a. Obtenga una ecuación de regresión estimada que sirva para estimar la cantidad en circulación de opciones otorgadas dada la cantidad de acciones comunes en circulación.
b. Emplee la ecuación de regresión estimada para estimar la cantidad en circulación de opciones otorgadas por una empresa que tiene 150 millones de acciones comunes en circulación.
c. ¿Cree que la ecuación de regresión estimada proporcione una buena predicción de la cantidad en circulación de opciones otorgadas? Emplee r^2 para justificar su respuesta.

60. El promedio industrial Dow Jones (DJIA) y el Estándar & Poor’s 500 (S & P) son índices que se emplean como una medida del movimiento general del mercado de valores. El DJIA se basa en los movimientos de los precios de 30 empresas grandes; el S&P 500 es un índice compuesto de 500 acciones. Algunos dicen que el S&P 500 es una mejor medida de la actividad del mercado de valores porque tiene una base más amplia. A continuación se presenta el precio de cierre del DJIA y del S&P 500 durante 20 semanas a partir del 9 de septiembre del 2005 (*Borron’s*, 30 de enero de 2006).



DJIAS&P500

Fecha	DJIA	S&P 500
9 de septiembre	10 679	1241
16 de septiembre	10 642	1238
23 de septiembre	10 420	1215
30 de septiembre	10 569	1229
7 de octubre	10 292	1196
14 de octubre	10 287	1187
21 de octubre	10 215	1180
28 de octubre	10 403	1198
4 de noviembre	10 531	1220
11 de noviembre	10 686	1235
18 de noviembre	10 766	1248
25 de noviembre	10 932	1268
2 de diciembre	10 878	1265
9 de diciembre	10 779	1259
16 de diciembre	10 876	1267
23 de diciembre	10 883	1269
30 de diciembre	10 718	1248
6 de enero	10 959	1285
13 de enero	10 960	1288
20 de enero	10 667	1261

a. Dé el diagrama de dispersión de estos datos empleando DJIA como variable independiente.
b. Obtenga la ecuación de regresión estimada.
c. Pruebe la significancia de la relación. Use $\alpha = 0.05$.
d. ¿Proporciona un buen ajuste la ecuación de regresión estimada? Explique.
e. Suponga que el precio de cierre del DJIA es 11 000. Estime el precio de cierre del S&P 500.
f. ¿Debe preocupar que el valor de 11 000 del DJIA empleado en el inciso e) para predecir el del S&P 500 se encuentre fuera del intervalo de los datos empleado para obtener la ecuación de regresión estimada?

61. Jensen Tire & Auto está por decidir si firma un contrato de mantenimiento para su nueva máquina de alineamiento y balanceo de neumáticos. Los gerentes piensan que los gastos de mantenimiento deberán estar relacionados con el uso y recolectan los datos siguientes sobre uso semanal (horas) y gastos anuales de mantenimiento (en cientos de dólares).



Jensen

Uso semanal (horas)	Gastos anuales de mantenimiento
13	17.0
10	22.0
20	30.0
28	37.0
32	47.0
17	30.5
24	32.5
31	39.0
40	51.5
38	40.0

a. Obtenga la ecuación de regresión estimada que relaciona gastos anuales de mantenimiento con el uso semanal.
b. Pruebe la significancia de la relación del inciso a) con 0.05 como nivel de significancia.
c. Jensen piensa que usará la nueva máquina 30 horas a la semana. Obtenga un intervalo de predicción de 95% para los gastos anuales de mantenimiento de la empresa.
d. Si el precio del contrato de mantenimiento es $3000 anuales, ¿recomendaría firmar el contrato de mantenimiento? ¿Por qué sí o por qué no?

62. En un determinado proceso de fabricación se cree que la velocidad (pies por minuto) de la línea de ensamblado afectaba al número de partes defectuosas halladas en el proceso de inspección. Para probar esto, los administradores idearon un procedimiento en el que la misma cantidad de partes por lote se examinaba visualmente a diferentes velocidades de la línea. Se recolectaron los datos siguientes.

Velocidad de la línea	Número de partes defectuosas halladas
20	21
20	19
40	15
30	16
60	14
40	17

a. Obtenga la ecuación de regresión estimada que relaciona velocidad de la línea de producción con el número de partes defectuosas encontradas.

b. Empleando el nivel de significancia 0.05, determine si la velocidad de la línea y el número de partes defectuosas halladas están relacionadas.
c. ¿Se ajusta bien a los datos la ecuación de regresión estimada?
d. Dé un intervalo de confianza de 95% para predecir el número medio de partes defectuosas si la velocidad de la línea es 50 pies por minuto.

63. Un hospital grande de una ciudad contrató a un sociólogo para que investigara la relación entre el número de días por año de ausencia con autorización, y la distancia (en millas) entre la casa y el trabajo del empleado. Se tomó una muestra de 10 empleados y se obtuvieron los datos siguientes.

archivo en CD
Absent

Distancia al trabajo	Número de días de ausencia
1	8
3	5
4	8
6	7
8	6
10	3
12	5
14	2
14	4
18	2

a. Elabore, con estos datos, un diagrama de dispersión.
b. Obtenga la ecuación de regresión de mínimos cuadrados.
c. ¿Existe una relación significativa entre las dos variables? Explique.
d. ¿Proporciona la ecuación de regresión estimada un buen ajuste? Explique
e. Emplee la ecuación de regresión estimada obtenida en el inciso b) para calcular un intervalo de confianza de 95% para el número esperado de ausencias (días) de los empleados que vivan a 5 millas de la empresa.

64. La autoridad de tránsito de una zona metropolitana importante desea determinar si hay relación entre la antigüedad de un autobús y los gastos de mantenimiento del mismo. En una muestra de 10 autobuses se obtuvieron los datos siguientes.



AgeCost

Antigüedad del autobús (años)	Costo de mantenimiento ($)
1	350
2	370
2	480
2	520
2	590
3	550
4	750
4	800
5	790
5	950

a. Empleando el método de mínimos cuadrados obtenga la ecuación de regresión estimada.
b. Haga una prueba para determinar si las dos variables están relacionadas de manera significativa con $\alpha = 0.05$.
c. ¿Proporciona la recta de mínimos cuadrados una buena aproximación a los datos observados? Explique.
d. Calcule un intervalo de predicción de 95% para los gastos de mantenimiento de un determinado autobús cuya antigüedad es de 4 años.

65. Un profesor de mercadotecnia de una universidad desea saber cuál es la relación entre las horas de estudio y la calificación en un curso. A continuación se presentan los datos obtenidos de 10 estudiantes que tomaron el curso el trimestre pasado.

archivo en CD
HoursPts

Horas de estudio	Calificación total
45	40
30	35
90	75
60	65
105	90
65	50
90	90
80	80
55	45
75	65

a. Obtenga la ecuación de regresión estimada que indica la relación entre calificación y horas de estudio.
b. Empleando $\alpha = 0.05$ pruebe la significancia del modelo.
c. Pronostique la calificación que obtendrá Mark Sweeney. Él estudió 95 horas.
d. Calcule un intervalo de predicción de 95% para la calificación de Mark Sweeney.

66. *Bloomberg Personal Finance* (julio/agosto 2001) publicó que la beta del mercado de Texas Instrument era 1.46. La beta del mercado de cada acción se determina mediante regresión lineal simple. En cada caso, la variable dependiente es la rentabilidad porcentual trimestral (revalorización más dividendos) menos el rendimiento porcentual que se hubiera obtenido en una inversión libre de riesgos (como tasa libre de riesgo se empleó la tasa Treasury Bill). La variable independiente es la rentabilidad porcentual trimestral (revalorización de capital más dividendos) para el mercado de valores (S&P 500) menos la rentabilidad porcentual de una inversión libre de riesgos. A partir de los datos trimestrales se desarrolla la ecuación de regresión estimada; la beta del mercado de la acción en cuestión es la pendiente de la ecuación de regresión estimada (b_1). La beta del mercado suele interpretarse como una medida de lo riesgoso de la acción. Si la beta del mercado es mayor a 1, la volatilidad de la acción es mayor al promedio en el mercado; si la beta del mercado es menor a 1, la volatilidad de la acción es menor al promedio en el mercado. Supóngase que las cifras siguientes son diferencias entre rentabilidad porcentual y rentabilidad libre de riesgos a lo largo de 10 trimestres de S&P 500 y Horizon Technology.



MktBeta

S&P 500	Horizon
1.2	−0.7
−2.5	−2.0
−3.0	−5.5
2.0	4.7
5.0	1.8
1.2	4.1
3.0	2.6
−1.0	2.0
0.5	−1.3
2.5	5.5

a. Obtenga la ecuación de regresión estimada que sirve para determinar la beta del mercado de Horizon Technology. ¿Cuál es la beta del mercado de Horizon Technology?
b. Empleando 0.05 como nivel de significancia, pruebe la significancia de la relación.
c. ¿Proporciona la ecuación de regresión estimada un buen ajuste? Explique.
d. Utilice las betas del mercado de Horizon Techology y de Texas Instrument para comparar los riesgos de estas dos acciones.

67. La Transactional Record Access Clearinghouse de la Universidad de Syracuse publica datos que muestran las posibilidades de una auditoría del Departamento de Tesorería de los Estados Unidos. En la tabla siguiente se muestra la media del ingreso bruto ajustado y el porcentaje de declaraciones que fueron auditadas en 20 municipios



IRSAudit

Municipio	Ingreso bruto ajustado	Porcentaje auditado
Los Ángeles	36 664	1.3
Sacramento	38 845	1.1
Atlanta	34 886	1.1
Boise	32 512	1.1
Dallas	34 531	1.0
Providence	35 995	1.0
San José	37 799	0.9
Cheyenne	33 876	0.9
Fargo	30 513	0.9
Nueva Orleans	30 174	0.9
Oklahoma City	30 060	0.8
Houston	37 153	0.8
Portland	34 918	0.7
Phoenix	33 291	0.7
Augusta	31 504	0.7
Albuquerque	29 199	0.6
Greensboro	33 072	0.6
Columbia	30 859	0.5
Nashville	32 566	0.5
Buffalo	34 296	0.5

a. Obtenga la ecuación de regresión estimada que sirve para pronosticar el porcentaje de auditorías dado un ingreso bruto ajustado.
b. Empleando como nivel de significancia 0.05, determine si hay relación entre el ingreso bruto ajustado y el porcentaje de auditorías.
c. ¿Proporciona la ecuación de regresión estimada un buen ajuste? Explique.
d. Emplee la ecuación de regresión estimada del inciso a) para calcular un intervalo de 95% de confianza para el porcentaje de auditorías en un municipio en el que el promedio del ingreso bruto ajustado es $35 000.

68. Una institución de un determinado país publicó evaluaciones sobre la satisfacción con el trabajo. Una de las cosas que se pedían en la encuesta era elegir (de una lista de factores) los cinco factores principales para la satisfacción en el trabajo. Después se pedía a los encuestados que indicaran su nivel de satisfacción con cada uno de esos cinco factores. En la tabla siguiente se presentan los porcentajes de personas para los que el factor indicado fue uno de los cinco factores principales, junto con una evaluación obtenida empleando el porcentaje de personas que consideraron al factor como uno de los principales y que estaban “muy satisfechos” o “satisfechos” con ese factor. (www.apse. gov.au/stateoftheservice).



JobSat

Factor	Cinco principales (%)	Evaluación (%)
Carga de trabajo adecuada	30	49
Posibilidad de ser creativo o de hacer innovaciones	38	64
Posibilidad de hacer contribuciones útiles a la sociedad	40	67
Obligaciones y expectativas claramente planteadas	40	69
Condiciones flexibles de trabajo	55	86
Buena relación de trabajo	60	85
Trabajo interesante	48	74
Oportunidad de hacer carrera	33	43
Oportunidad de desarrollar sus habilidades	46	66
Oportunidad de utilizar sus habilidades	50	70
Retroalimentación y reconocimiento al esfuerzo realizado	42	53
Salario	47	62
Poder ver resultados tangibles del trabajo	42	69

a. Elabore un diagrama de dispersión colocando en el eje horizontal los porcentajes de los factores principales y en el eje vertical la evaluación correspondiente.
b. ¿Qué indica, respecto a la relación entre las dos variables, el diagrama de dispersión elaborado en el inciso a)?
c. Obtenga la ecuación de regresión estimada que sirva para pronosticar la evaluación (%) dado el porcentaje del factor (%).
d. Empleando como nivel de significancia 0.05 realice una prueba para determinar la significancia de la relación.
e. ¿Proporciona la ecuación de regresión estimada un buen ajuste?
f. Dé el valor del coeficiente de correlación muestral.

Caso problema 1 Medición del riesgo en el mercado bursátil

Una medida del riesgo o volatilidad de una acción es la desviación estándar del rendimiento durante un lapso de tiempo. Aunque la desviación estándar es fácil de calcular, no toma en cuenta la variación del precio de una acción en función de un índice estándar del mercado, como el S&P 500. Por esta razón, muchos analistas financieros prefieren emplear otra medida, conocida como *beta,* para medir el riesgo.



Beta

La beta de una acción se determina mediante regresión lineal simple. La variable independiente es la rentabilidad total de la acción de que se trate y la variable independiente es la rentabilidad total del mercado de valores.* En este caso problema se usará el índice S&P 500 como medida de la rentabilidad total del mercado de valores y se obtendrá una ecuación de regresión estimada usando datos mensuales. La beta de una acción es la pendiente en la ecuación de regresión estimada (b_1). En el archivo Beta del disco compacto que se distribuye con el libro se proporciona la rentabilidad total de ocho acciones comunes muy conocidas y la del S&P 500 a lo largo de 36 meses.

El valor beta del mercado de valores siempre será 1; por lo tanto, una acción que tienda a subir o a bajar con el mercado de valores tendrá también una beta cercana a 1. Betas mayores a 1 corresponden a acciones que son más volátiles que el mercado y betas menores a 1 corresponden a acciones menos volátiles que el mercado. Por ejemplo, si la beta de una acción es 1.4, esta acción es 40% *más* volátil que el mercado, y si la beta de una acción es 0.4, la acción es 60% menos volátil que el mercado.

*Diversas fuentes emplean diferentes métodos para calcular las betas. Por ejemplo, algunas fuentes, antes de calcular la ecuación de regresión estimada, restan, de la variable independiente, la rentabilidad que podría haberse obtenido con una inversión libre de riesgos [por ejemplo, letras del tesoro (Estados Unidos)(T-bills)]. Otras, para la rentabilidad total del mercado de valores emplean diversos índices; por ejemplo, *Value line* calcula las betas usando el índice compuesto de la bolsa de Nueva York

Reporte administrativo

Se le ha encomendado la tarea de analizar las características del riesgo de estas acciones. Elabore un informe que comprenda los puntos siguientes, sin limitarse sólo a ellos.

a. Calcular los estadísticos descriptivos de cada una de las acciones y del S&P 500. Hacer comentarios sobre los resultados. ¿Qué acción es la más volátil?
b. Calcular la beta de cada acción. ¿Cuál de estas acciones se esperaría que se comportara mejor en un mercado de alta calidad? ¿Cuál conservaría mejor su valor en un mercado para el sector popular?
c. Haga un comentario sobre qué tanto de la rentabilidad de cada una de las acciones es explicado por el mercado.

Caso problema 2 Departamento de Transporte de Estados Unidos

Como parte de un estudio sobre seguridad en el transporte, el Departamento de Transporte de Estados Unidos, de una muestra de 42 ciudades, recogió datos sobre el número de accidentes fatales por cada 1000 licencias y sobre el porcentaje de licencia de conductores menores de 21 años. A continuación se presentan los datos recogidos en el lapso de un año. Estos datos se encuentran también en el archivo titulado Safety del disco compacto que se distribuye con el libro.



Safety

Porcentaje de menores de 21 años	Accidentes fatales por 1000 licencias	Porcentaje de menores de 21 años	Accidentes fatales por 1000 licencias
13	2.962	17	4.100
12	0.708	8	2.190
8	0.885	16	3.623
12	1.652	15	2.623
11	2.091	9	0.835
17	2.627	8	0.820
18	3.830	14	2.890
8	0.368	8	1.267
13	1.142	15	3.224
8	0.645	10	1.014
9	1.028	10	0.493
16	2.801	14	1.443
12	1.405	18	3.614
9	1.433	10	1.926
10	0.039	14	1.643
9	0.338	16	2.943
11	1.849	12	1.913
12	2.246	15	2.814
14	2.855	13	2.634
14	2.352	9	0.926
11	1.294	17	3.256

Informe administrativo

1. Presente resúmenes numéricos y gráficos de los datos.
2. Emplee el análisis de regresión para investigar la relación entre el número de accidentes fatales y el porcentaje de conductores menores de 21 años. Analice sus hallazgos.
3. ¿Qué conclusión y qué recomendaciones puede deducir de su análisis?

Caso problema 3 Donaciones de los ex alumnos

Las donaciones de los ex alumnos son una importante fuente de ingresos para las universidades. Si los gerentes pudieran determinar los factores que influyen sobre el aumento del porcentaje de alumnos que hace donaciones, podrían poner en marcha políticas que llevarían a ganancias mayores. Las investigaciones indican que estudiantes más satisfechos con la relación con sus profesores tienen más probabilidad de titularse, lo que a su vez puede llevar al aumento del porcentaje de alumnos que haga donaciones. En la tabla 14.13 se muestran datos de 48 universidades de Estados Unidos (*American's Best Collage,* edición del año 2000). La columna titulada "% de grupos con menos de 20" muestra el porcentaje de grupos con menos de 20 alumnos. La columna que tiene como título "Tasa de estudiantes/facultad" da el número de estudiantes inscritos, dividido entre el número total de facultades. Por último, la columna que tiene como título "Tasa de alumnos que donan" da el porcentaje de alumnos que han hecho alguna donación a la universidad.

Reporte administrativo

1. Presente resúmenes numéricos y gráficos de los datos.
2. Emplee el análisis de regresión para obtener una ecuación de regresión estimada que sirva para pronosticar el porcentaje de los estudiantes que hacen donaciones dado el porcentaje de grupos con menos de 20 estudiantes.
3. Use el análisis de regresión para obtener una ecuación de regresión estimada que sirva para pronosticar el porcentaje de los alumnos que hacen donaciones dada la proporción de estudiantes por facultad.
4. ¿Cuál de las dos ecuaciones de regresión estimada muestra un mejor ajuste? Con esa ecuación de regresión estimada realice un análisis de residuales y discuta sus hallazgos y conclusiones.
5. ¿Qué conclusiones y recomendaciones puede obtener de este análisis?

Caso problema 4 Valor de los equipos de béisbol de la liga mayor

Un grupo encabezado por John Henry pagó $700 millones por la adquisición del equipo Boston Red Sox (Medias Rojas de Boston) en 2002, aun cuando el Boston Red Sox no había ganado la serie mundial desde 1918 y tenía una pérdida de operación de $11.4 millones de 2001. Es más, la revista *Forbes* estima que el valor actual del equipo es en realidad $426 millones. *Forbes* atribuye la diferencia entre valor actual del equipo y precio que los inversionistas están dispuestos a pagar, al hecho de que la compra de un equipo suele incluir la adquisición de una red de cable exageradamente subvaluada. Por ejemplo, con la compra del equipo, los nuevos propietarios obtuvieron también la New England Sports Network. En la tabla 14.14 se presentan los datos de 30 equipos de la liga mayor (*Forbes*, 15 de abril de 2002). En la columna titulada Valor se da el valor de los equipos con base en las actuales negociaciones con los estadios, sin deducción de deudas. En la columna titulada Ingreso se presentan las ganancias sin intereses, impuestos y depreciación.

Informe administrativo

1. Presente resúmenes numéricos y gráficos de los datos.
2. Use el análisis de regresión para investigar la relación entre valor e ingreso. Discuta sus hallazgos.
3. Use el análisis de regresión para investigar la relación entre valor y ganancias. Discuta sus hallazgos.
4. ¿Qué conclusiones y recomendaciones puede sacar de este análisis?



TABLA 14.13 DATOS DE 48 UNIVERSIDADES NACIONALES

	% de grupos con menos de 20	Tasa de estudiantes/ facultad	Tasa de alumnos que donan
Boston College	39	13	25
Brandeis University	68	8	33
Brown University	60	8	40
California Institute of Technology	65	3	46
Carnegie Mellon University	67	10	28
Case Western Reserve Univ.	52	8	31
College of William and Mary	45	12	27
Columbia University	69	7	31
Cornell University	72	13	35
Dartmouth College	61	10	53
Duke University	68	8	45
Emory University	65	7	37
Georgetown University	54	10	29
Harvard University	73	8	46
Johns Hopkins University	64	9	27
Lehigh University	55	11	40
Massachusetts Inst. of Technology	65	6	44
New York University	63	13	13
Northwestern University	66	8	30
Pennsylvania State Univ.	32	19	21
Princeton University	68	5	67
Rice University	62	8	40
Stanford University	69	7	34
Tufts University	67	9	29
Tulane University	56	12	17
U. of California–Berkeley	58	17	18
U. of California–Davis	32	19	7
U. of California–Irvine	42	20	9
U. of California–Los Angeles	41	18	13
U. of California–San Diego	48	19	8
U. of California–Santa Barbara	45	20	12
U. of Chicago	65	4	36
U. of Florida	31	23	19
U. of Illinois–Urbana Champaign	29	15	23
U. of Michigan–Ann Arbor	51	15	13
U. of North Carolina–Chapel Hill	40	16	26
U. of Notre Dame	53	13	49
U. of Pennsylvania	65	7	41
U. of Rochester	63	10	23
U. of Southern California	53	13	22
U. of Texas–Austin	39	21	13
U. of Virginia	44	13	28
U. of Washington	37	12	12
U. of Wisconsin–Madison	37	13	13
Vanderbilt University	68	9	31
Wake Forest University	59	11	38
Washington University–St. Louis	73	7	33
Yale University	77	7	50

TABLA 14.14 DATOS DE LOS EQUIPOS DE LA LIGA MAYOR DE BASQUETBOL



MLB

Equipo	Valor	Ganancia	Ingreso
New York Yankees	730	215	18.7
New York Mets	482	169	14.3
Los Angeles Dodgers	435	143	−29.6
Boston Red Sox	426	152	−11.4
Atlanta Braves	424	160	9.5
Seattle Mariners	373	166	14.1
Cleveland Indians	360	150	−3.6
Texas Rangers	356	134	−6.5
San Francisco Giants	355	142	16.8
Colorado Rockies	347	129	6.7
Houston Astros	337	125	4.1
Baltimore Orioles	319	133	3.2
Chicago Cubs	287	131	7.9
Arizona Diamondbacks	280	127	−3.9
St. Louis Cardinals	271	123	−5.1
Detroit Tigers	262	114	12.3
Pittsburgh Pirates	242	108	9.5
Milwaukee Brewers	238	108	18.8
Philadelphia Phillies	231	94	2.6
Chicago White Sox	223	101	−3.8
San Diego Padres	207	92	5.7
Cincinnati Reds	204	87	4.3
Anaheim Angels	195	103	5.7
Toronto Blue Jays	182	91	−20.6
Oakland Athletics	157	90	6.8
Kansas City Royals	152	85	2.2
Tampa Bay Devil Rays	142	92	−6.1
Florida Marlins	137	81	1.4
Minnesota Twins	127	75	3.6
Montreal Expos	108	63	−3.4

Apéndice 14.1 Deducción de la fórmula de mínimos cuadrados empleando el cálculo

Como ya se indicó en este capítulo, el método de mínimos cuadrados se usa para determinar los valores de b_0 y b_1 que minimicen la suma de los cuadrados de los residuales. La suma de los cuadrados de los residuales está dada por

$$\Sigma(y_i - \hat{y}_i)^2$$

Sustituyendo $\hat{y}_i = b_0 + b_1x_i$, se obtiene

$$\Sigma(y_i - b_0 - b_1x_i)^2 \tag{14.34}$$

como expresión que hay que minimizar.

Para minimizar la expresión (14.14), se sacan las derivadas parciales respecto a b_0 y b_1, se igualan a cero y despeja. Haciendo esto se obtiene

$$\frac{\partial\Sigma(y_i - b_0 - b_1x_i)^2}{\partial b_0} = -2\Sigma(y_i - b_0 - b_1x_i) = 0 \quad \textbf{(14.35)}$$

$$\frac{\partial\Sigma(y_i - b_0 - b_1x_i)^2}{\partial b_1} = -2\Sigma x_i(y_i - b_0 - b_1x_i) = 0 \quad \textbf{(14.36)}$$

Dividiendo la ecuación (14.35) entre dos y haciendo las sumas por separado, se obtiene

$$-\Sigma y_i + \Sigma b_0 + \Sigma b_1x_i = 0$$

Llevando Σy_i al otro lado del signo igual y observando que $\Sigma b_0 = nb_0$, se obtiene

$$nb_0 + (\Sigma x_i)b_1 = \Sigma y_i \quad \textbf{(14.37)}$$

Simplificaciones algebraicas similares aplicadas a la ecuación (14.36) producen

$$(\Sigma x_i)b_0 + (\Sigma x_i^2)b_1 = \Sigma x_iy_i \quad \textbf{(14.38)}$$

A las ecuaciones (14.37) y (14.38) se les conoce como *ecuaciones normales*. Despejando b_0 en la ecuación (14.37) se obtiene

$$b_0 = \frac{\Sigma y_i}{n} - b_1\frac{\Sigma x_i}{n} \quad \textbf{(14.39)}$$

Usando la ecuación (14.39) para sustituir a b_0 en la ecuación (14.38) da

$$\frac{\Sigma x_i\Sigma y_i}{n} - \frac{(\Sigma x_i)^2}{n}b_1 + (\Sigma x_i^2)b_1 = \Sigma x_iy_i \quad \textbf{(14.40)}$$

Reordenando los términos de la ecuación (14.40), se obtiene

$$b_1 = \frac{\Sigma x_iy_i - (\Sigma x_i\Sigma y_i)/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \quad \textbf{(14.41)}$$

Como $\bar{y} = \Sigma y_i/n$ y $\bar{x} = \Sigma x_i/n,$ la ecuación (14.39) se puede rescribir como

$$b_0 = \bar{y} - b_1\bar{x} \quad \textbf{(14.42)}$$

Las ecuaciones (14.41) y (14.42) son las fórmulas (14.6) y (14.7) usadas en este capítulo para calcular los coeficientes de la ecuación de regresión estimada.

Apéndice 14.2 Una prueba de significancia usando correlación

Empleando el coeficiente de correlación muestral r_{xy}, también se puede determinar si la relación lineal entre x y y es significativa mediante la siguiente prueba de hipótesis acerca del coeficiente de correlación muestral.

$$H_0\text{: } \rho_{xy} = 0$$
$$H_a\text{: } \rho_{xy} \neq 0$$

Si H_0 es rechazada, se concluye que el coeficiente de correlación no es igual a cero y que la relación entre las dos variables no es significativa. A continuación se presenta esta prueba de significancia.

PRUEBA DE SIGNIFICANCIA USANDO CORRELACIÓN

$$H_0: \rho_{xy} = 0$$
$$H_a: \rho_{xy} \neq 0$$

ESTADÍSTICO DE PRUEBA

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}} \tag{14.43}$$

REGLA DE RECHAZO

Método del valor-p: Rechazar H_0 si valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $t \leq -t_{\alpha/2}$ o si $t \geq t_{\alpha/2}$

donde $t_{\alpha/2}$ pertenece a la distribución t con $n - 2$ grados de libertad.

En la sección 14.4 con una muestra $n = 10$ se encontró que el coeficiente de correlación muestral para la población de estudiantes y las ventas trimestrales era $r_{xy} = 0.9501$. El estadístico de prueba es

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}} = 0.9501\sqrt{\frac{10-2}{1-(0.9501)^2}} = 8.61$$

En la tabla de la distribución t se encuentra que para $n - 2 = 10 - 2 = 8$ grados de libertad, $t = 3.355$ proporciona un área de 0.005 en la cola superior. Por lo tanto, al área en la cola superior que corresponde al estadístico de prueba $t = 8.61$ debe ser menor a 0.005. Como esta prueba es una prueba de dos colas, se duplica este valor y se concluye que el valor t que corresponde a $t = 8.62$ debe ser menor a $2(0.005) = 0.01$. Con Excel o con Minitab se obtiene valor-p = 0.000. Como el valor-p es menor a $\alpha = 0.01$, se rechaza H_0 y se concluye que r_{xy} no es igual a cero. Esta evidencia es suficiente para concluir que entre la población de estudiantes y las ventas trimestrales existe una relación lineal significativa.

Obsérvese que el valor del estadístico de prueba t y la conclusión sobre la significancia de la relación son idénticos con los resultados obtenidos en la prueba t de la sección 14.5, en donde se usó la ecuación de regresión estimada $\hat{y} = 60 + 5x$. El análisis de regresión permite obtener una conclusión sobre la relación entre las variables x y y; además, permite obtener la ecuación que indica cuál es la relación entre las variables. Por consiguiente, la mayoría de los analistas emplean paquetes modernos de software para realizar el análisis de regresión y encuentran que el empleo de la correlación como prueba de significancia es innecesario.

Apéndice 14.3 Análisis de regresión con Minitab



Armand's

En la sección 14.7 mostrando los resultados que da Minitab para el problema de Armand's Pizza Parlors se estudió la solución de los problemas de regresión mediante el empleo de paquetes de software. En este apéndice se describen los pasos necesarios al emplear Minitab para generar esos resultados. Primero, en una hoja de cálculo de Minitab se ingresan los datos. Los datos de las poblaciones de estudiantes se ingresan en la columna C1 y los datos de las ventas trimestrales se ingresan en la columna C2. Los nombres de las variables Pop y Sales (Ventas) se ingresan como encabezados de esas columnas. En la descripción de los pasos a seguir, para referirse a los datos se emplearán los nombres de las variables o los indicadores de las columnas C1 y C2. Los

pasos siguientes describen cómo usar Minitab para obtener los resultados del análisis de regresión que se muestran en la figura 14.10.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Seleccionar el menú **Regression**
Paso 3. Elegir **Regression**
Paso 4. Cuando aparezca el cuadro de diálogo Regresión:
Ingresar Sales en el cuadro **Response**
Ingresar Pop en el cuadro **Predictors**
Clic en el botón **Options**
Cuando aparezca el cuadro de diálogo Regression-Options:
Ingresar 10 en el cuadro **Prediction intervals for new observations**
Clic en **OK**
Cuando aparezca el cuadro de diálogo Regression:
Clic en **OK**

El cuadro de diálogo de Minitab tiene otras posibilidades más que se pueden aprovechar seleccionando las opciones deseadas. Por ejemplo, para obtener una gráfica de residuales, en la que los valores pronosticados $\hat{y}$ aparezcan en el eje horizontal y los valores de los residuales estandarizados en el eje vertical, el paso 4 deberá ser como sigue:

Paso 4 Cuando aparezca el cuadro de diálogo Regression:
Ingresar Sales en el cuadro **Response**
Ingresar Pop en el cuadro **Predictors**
Clic en el botón **Graphs**
Cuando aparezca el cuadro de diálogo Regression-Graphs:
Seleccionar **Standardized** en Residuals for Plots
Seleccionar **Residuals versus fits** en Residual Plots
Clic en **OK**
Cuando aparezca el cuadro de diálogo Regression:
Clic en **OK**

Apéndice 14.4 Análisis de regresión con Excel



En este apéndice se ilustra el uso de la herramienta de Excel para realizar los cálculos del análisis de regresión empleando el problema de Armand´s Pizza Parlors. Consúltese la figura 14.23, para seguir la descripción de los pasos. En las celdas A1:C1 de la hoja de cálculo se ingresan los rótulos Restaurante, Población y Ventas. Para identificar cada una de las 10 observaciones, se ingresan los números del 1 al 10 en las celdas A2:A11. Los datos muestrales se ingresan en las celdas B2:C11. Los pasos siguientes indican cómo obtener los resultados del análisis de regresión.

Paso 1. Seleccionar el menú **Herramientas**
Paso 2. Elegir el menú **Análisis de datos**
Paso 3. Elegir **Regresión** en el menú de Funciones para análisis
Paso 4. Clic en **OK**
Paso 5. Cuando aparezca el cuadro de diálogo Regresión:
Ingresar C1:C11 en el cuadro **Rango Y de entrada**
Ingresar B1:B11 en el cuadro **Rango X de entrada**
Seleccionar **Rótulos**
Seleccionar **Nivel de confianza**
Ingresar **99** en el cuadro **Nivel de confianza**
Seleccionar **Rango de salida**
Ingresar A13 en el cuadro **Rango de salida**
(También se puede ingresar cualquier celda, de la esquina superior izquierda, para indicar dónde deberán empezar los resultados.)
Clic en **OK**

FIGURA 14.23 SOLUCIÓN CON EXCEL AL PROBLEMA DE ARMAND'S PIZZA PARLORS

	A	B	C	D	E	F	G	H	I	J
1	**Restaurant**	**Population**	**Sales**							
2	1	2	58							
3	2	6	105							
4	3	8	88							
5	4	8	118							
6	5	12	117							
7	6	16	137							
8	7	20	157							
9	8	20	169							
10	9	22	149							
11	10	26	202							
12										
13	SUMMARY OUTPUT									
14										
15	*Regression Statistics*									
16	Multiple R	0.9501								
17	R Square	0.9027								
18	Adjusted R Square	0.8906								
19	Standard Error	13.8293								
20	Observations	10								
21										
22	ANOVA									
23		*df*	*SS*	*MS*	*F*	*Significance F*				
24	Regression	1	14200	14200	74.2484	2.55E-05				
25	Residual	8	1530	191.25						
26	Total	9	15730							
27										
28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*	
29	Intercept	60	9.2260	6.5033	0.0002	38.7247	81.2753	29.0431	90.9569	
30	Population	5	0.5803	8.6167	2.55E-05	3.6619	6.3381	3.0530	6.9470	
31										
32										
33										
34										

La primera sección de los resultados, titulada *Estadísticas de regresión,* contiene resúmenes estadísticos como el coeficiente de determinación (R*2). La segunda sección de los resultados, titulada Análisis de varianza, contiene la tabla del análisis de varianza. La última sección de los resultados, que no tiene ningún título, contiene los coeficientes de regresión estimados e información relacionada con ellos. A continuación se da la interpretación de los resultados de la regresión empezando con la información contenida en las celdas A28:I30

Interpretación de los resultados de la ecuación de regresión estimada

La intersección de la recta de regresión con el eje y, $b_0 = 60$, aparece en la celda B29 y la pendiente de la recta de regresión estimada, $b_1 = 5$, aparece en la celda B30. El rótulo Intercepción en la celda A29 y el rótulo Población en la celda A30 sirven para identificar estos dos valores.

En la sección 14.5 se mostró que la desviación estándar estimada de b_1 es $s_{b_1} = 0.5803$. Obsérvese que el valor de la celda C30 es 0.5803. El rótulo Error típico que aparece en la celda C28, es la manera en que Excel indica que el valor de la celda C30 es el error estándar o la desviación estándar de b_1. Recuérdese que en la prueba t de significancia de la relación fue necesario calcular el estadístico t, $t = b_1/s_{b_1}$. Empleando los datos de Armand's, se obtuvo como valor t, $t = 5/.5803 = 8.62$. El rótulo *Estadístico t* de la celda D28 sirve para recordar que en la celda D30 se encuentra el valor del estadístico t.

El valor en la celda E30 es el valor –p que corresponde a la prueba t de significancia. El valor-p que da Excel en la celda E30, está en notación científica. Para obtener este valor en notación decimal, se recorre el punto decimal 5 lugares a la izquierda, con lo que se obtiene 0.0000255. Dado que valor-$p = 0.0000255 < \alpha = 0.01$, se rechaza H_0 y se concluye que entre la población de estudiantes y las ventas trimestrales existe una relación significativa.

La información de las celdas F28:I30 se emplea para obtener estimaciones por Intervalos de confianza para la intersección con el eje y y la pendiente de la ecuación de regresión estimada. Excel siempre da los límites inferior y superior de un intervalo de 95% de confianza. Como en el paso 4 se seleccionó Intervalo de confianza y se ingresó 99 en el cuadro de Nivel de Confianza, la herramienta de Excel para regresión da también los límites inferior y superior de un intervalo de 99% de confianza. El valor en la celda H30 es el límite inferior de la estimación por intervalo del 99% de confianza de b_1 y el valor en la celda I30 es el límite superior. Por lo tanto, una vez redondeada, el intervalo de 99% de confianza para estimar b_1 va de 3.05 a 6.95. Los valores en las celdas F30 a G30 proporcionan los límites inferior y superior del intervalo de 95% de confianza. El intervalo de 95% de confianza va de 3.66 a 6.34.

Interpretación de los resultados del ANOVA

La información en las celdas A22:F26 es un resumen de los cálculos del análisis de varianza. Las tres fuentes de variación están rotuladas como Regresión, Residuo y Total. La etiqueta *df* en la celda B23 representa los grados de libertad, la etiqueta *SS* en la celda C23 representa la suma de los cuadrados y la etiqueta *MS* en la celda D23 representa el cuadrado de la media.

En la sección 14.5 se dijo que el error cuadrado medio, que se obtiene dividiendo el error o la suma de cuadrados del residual entre sus grados de libertad, proporciona una estimación de σ^2. El valor en la celda D25, 191.25, es el error cuadrado medio de los resultados de regresión para el problema de Armand's. En la sección 14.5 se mostró que también se puede usar una prueba F como prueba de significancia en la regresión. El valor en la celda F24, 0.0000255, es el valor-p que corresponde a la prueba F de significancia. Dado que valor-$p = 0.0000255 < \alpha = 0.01$, se rechaza H_0 y se concluye que se tiene una relación significante entre la población de estudiantes y las ventas trimestrales. En la celda F23, el rótulo que emplea Excel para identificar el valor-p de la prueba F de significancia es *Valor crítico de F.*

El rótulo Valor crítico de F se entiende mejor si se considera el valor en la celda F24 como el nivel de significancia observado en la prueba F.

Interpretación de los estadísticos de regresión de los resultados

El coeficiente de determinación, 0.9027, aparece en la celda B17; el rótulo correspondiente, Coeficiente de determinación R*2, aparece en la celda A17. La raíz cuadrada del coeficiente de determinación es el coeficiente de correlación muestral, 0.9501, que aparece en la celda B16. Obsérvese que para identificar este valor, Excel emplea como rótulo Coeficiente de correlación múltiple. En la celda A19, el rótulo Error Estándar se usa para identificar el valor del error estándar de estimación que aparece en la celda B19. Así que el error estándar de estimación es 13.8293. Hay que tener presente que en los resultados de Excel, el rótulo Error típico aparece en dos lugares. En la sección de los resultados titulada Estadísticas de regresión, el rótulo Error típico se refiere a la estimación de σ; en la sección de los resultados correspondiente a la Ecuación de regresión estimada, el rótulo *Error típico* se refiere a s_{b_1}, la desviación estándar de la distribución muestral de b_1.



CAPÍTULO 15

Regresión múltiple

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

INTERNATIONAL PAPER*
PURCHASE, NUEVA YORK

International Paper es la mayor empresa del mundo que se dedica a la producción de papel y productos forestales. Esta empresa da empleo a más de 117 000 personas en casi 50 países y exporta sus productos a más de 130 naciones. International Paper produce materiales de construcción como madera para construcción y madera de contrachapa; materiales de empaque como vasos y recipientes desechables; materiales para empaque industrial como cajas de cartón corrugado y embalaje de expedición, además de una gran variedad de papeles para fotocopiadoras, impresoras, libros y material para publicidad.

En la fabricación de los productos de papel, se procesan virutas de madera y productos químicos en molinos de pulpa para obtener la pulpa de madera. Después, la pulpa de madera se emplea para producir los productos de papel. Para los productos de papel blanco, es necesario blanquear la pulpa con objeto de eliminar cualquier alteración cromática. El agente blanqueador esencial en este proceso es el dióxido de cloro, el cual, debido a su naturaleza combustible, suele producirse en un molino de pulpa, de donde, en forma de solución, es transportado a través de una tubería a la torre de blanqueo del molino de pulpa. Con objeto de mejorar uno de los procesos empleados en la producción del dióxido de cloro, se estudió el control y la eficiencia del proceso. Uno de los aspectos estudiados fue la velocidad de alimentación de las sustancias químicas que intervienen en la producción del dióxido de cloro.

En la producción del dióxido de cloro intervienen cuatro sustancias químicas que llegan, a velocidades controladas, al generador de dióxido de cloro. El dióxido de cloro producido en el generador se recibe en un absorbente, de donde el dióxido de cloro gaseoso es absorbido en agua helada formando una solución de dióxido de cloro. A continuación la solución pasa a un molino de papel. Parte esencial del control de este proceso es la velocidad de alimentación de las sustancias químicas. Antes, los operadores fijaban la velocidad de alimentación de las sustancias químicas, método que llevaba a un sobrecontrol del proceso. Debido a esto, los ingenieros químicos encargados del molino solicitaron que, como ayuda para fijar las velocidades de alimentación de estas sustancias, se obtuviera un conjunto de ecuaciones de control, una para la alimentación de cada una de las sustancias químicas.



El análisis de regresión múltiple se empleó para obtener un mejor proceso de blanqueo en la fabricación de productos de papel blanco. © Lester Lefkowitz/Corbis.

Empleando el análisis de regresión múltiple, los analistas obtuvieron, para cada una de las cuatro sustancias químicas empleadas en el proceso, una ecuación de regresión múltiple estimada. Cada ecuación relaciona la producción de dióxido de cloro con la cantidad de la sustancia química empleada y con la concentración de la solución de dióxido de cloro. En cada uno de los molinos se programó en una microcomputadora el conjunto de las cuatro ecuaciones obtenidas. Con el nuevo sistema, los operadores ingresan al sistema la concentración de la solución de dióxido de cloro y la velocidad de producción deseadas; el paquete de software calcula la velocidad de alimentación de la sustancia química que permite obtener esa velocidad de producción deseada. Desde que los operadores empezaron a usar las ecuaciones de control, aumentó significativamente la eficiencia del generador de dióxido de cloro así como la cantidad de veces en las que la concentración de cloro caía dentro del rango aceptable.

Este ejemplo muestra el empleo del análisis de regresión múltiple en la obtención de mejores procesos de blanqueo para producir productos de papel blanco. En este capítulo se verá el uso de los paquetes de software para tales propósitos. La mayor parte de los conceptos presentados en el capítulo 14, para la regresión lineal simple, pueden extenderse a la regresión múltiple.

---

* Los autores agradecen a Mariam Williams y Hill Griggs por proporcionar este artículo para *La estadística en la práctica*. Esta aplicación fue elaborada por Champion International Corporation, empresa que en 2000 se volvió parte de International Paper.

En el capítulo 14 se presentó la regresión lineal simple y se mostró su uso en la obtención de una ecuación de regresión estimada que describe la relación entre dos variables. Recuérdese que la variable que se predice o explica es la variable dependiente y la variable que se usa para predecir o explicar la variable dependiente es la variable independiente. En este capítulo se continúa con el estudio del análisis de regresión considerando, ahora, las situaciones en las que intervienen dos o más variables independientes. Este estudio, al que se le conoce como **análisis de regresión múltiple**, permite tomar más factores en consideración y obtener estimaciones mejores que las que son posibles con la regresión lineal simple.

15.1 Modelo de regresión múltiple

El análisis de regresión múltiple estudia la relación de una variable dependiente con dos o más variables independientes. Para denotar el número de variables independientes se suele usar p.

Modelo de regresión y ecuación de regresión

Los conceptos de modelo de regresión y ecuación de regresión vistos en el capítulo previo, son aplicables en el caso de la regresión múltiple. A la ecuación que describe cómo está relacionada la variable dependiente y con las variables independientes $x_1, x_2, \ldots, x_p$ se le conoce como **modelo de regresión múltiple**. Se supone que el modelo de regresión múltiple toma la forma siguiente

MODELO DE REGRESIÓN MÚLTIPLE

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \quad \textbf{(15.1)}$$

En el modelo de regresión múltiple, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, son parámetros y el término del error ϵ (la letra griega épsilon) es una variable aleatoria. Examinando con atención este modelo se ve que y es una función lineal de $x_1, x_2, \ldots, x_p$ (la parte $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$) más el término del error ϵ. El término del error corresponde a la variabilidad en y que no puede atribuirse o explicarse al efecto lineal de las p variables independientes.

En la sección 15.4 se discutirán los supuestos para el modelo de regresión múltiple y para ϵ. Uno de los supuestos es que la media o valor esperado de ϵ es cero. Una consecuencia de este supuesto es que la media o valor esperado de y, que se denota $E(y)$, es igual a $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. A la ecuación que describe cómo está relacionada la media de y con $x_1, x_2, \ldots, x_p$ se le conoce como **ecuación de regresión múltiple**.

ECUACIÓN DE REGRESIÓN MÚLTIPLE

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad \textbf{(15.2)}$$

Ecuación de regresión múltiple estimada

Si se conocieran los valores de $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, se podría usar la ecuación (15.2) para calcular la media de las y para valores dados de $x_1, x_2, \ldots, x_p$. Desafortudamente, los valores de estos parámetros no suelen conocerse, es necesario estimarlos a partir de datos muestrales. Para calcular los valores de los estadísticos muestrales $b_1, b_2, \ldots, b_p$, que se usan como estimadores puntuales de los parámetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ se emplea una muestra aleatoria simple. Con los estadísticos muestrales se obtiene la siguiente **ecuación de regresión múltiple estimada**.

ECUACIÓN DE REGRESIÓN MÚLTIPLE ESTIMADA

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p \quad \textbf{(15.3)}$$

donde

$b_0, b_1, b_2, \ldots, b_p$ son las estimaciones de $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$
$\hat{y}$ = valor estimado de la variable dependiente

Este proceso de estimación en la regresión múltiple se muestra en la figura 15.1.



15.2 Método de mínimos cuadrados

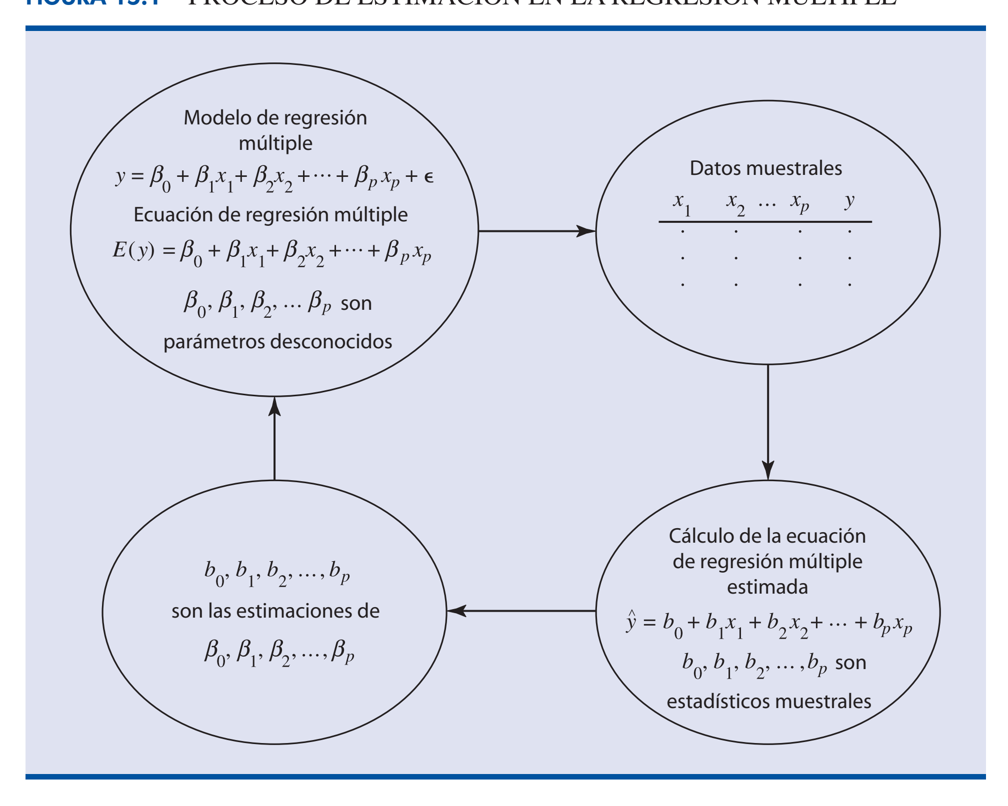
En el capítulo 14, se usó el **método de mínimos cuadrados** para obtener la ecuación de regresión estimada que permitía aproximar mejor la relación lineal entre las variables dependiente e independiente. Este método también se usa para obtener la ecuación de regresión múltiple estimada. El criterio en el método de mínimos cuadrados, como ya se dijo, es el siguiente.

CRITERIO DE MÍNIMOS CUADRADOS

$$\min \Sigma(y_i - \hat{y}_i)^2 \quad \textbf{(15.4)}$$

En la regresión lineal simple, b_0 y b_1 son los estadísticos muestrales que se usan para estimar los parámetros β_0 y β_1. En la regresión múltiple, en el proceso de inferencia estadística análogo, b_0, b_1, b_2, ..., b_p denotan los estadísticos muestrales que se usan para estimar los parámetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

FIGURA 15.1 PROCESO DE ESTIMACIÓN EN LA REGRESIÓN MÚLTIPLE



donde

y_i = valor observado en la variable dependiente en la observación i
$\hat{y}_i$ = valor estimado para la variable dependiente en la observación i

Los valores estimados de la variable dependiente se calculan empleando la ecuación de regresión múltiple estimada

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

Como indica la expresión (15.4), el método de mínimos cuadrados emplea datos muestrales para obtener los valores de b_0, b_1, b_2, . . ., b_p que hacen que la suma de los cuadrados de los residuales [las diferencias entre los valores observados de la variable dependiente (y_i) y los valores estimados de la variable dependiente ($\hat{y}_i$)] sea un mínimo.

En el capítulo 14 se dieron las fórmulas para calcular los estimadores b_0 y b_1 para la ecuación de regresión lineal simple estimada $\hat{y} = b_0 + b_1x$ empleando el método de mínimos cuadrados. Con conjuntos de datos relativamente pequeños, fue posible usar esas fórmulas para obtener b_0 y b_1 mediante cálculos manuales. En la regresión múltiple, en cambio, las fórmulas para calcular b_0, b_1, b_2, . . . , b_p emplean álgebra de matrices y quedan fuera del alcance de este texto. Por esta razón, en el estudio de la regresión múltiple, se centrará la atención en el uso de los paquetes de software para obtener la ecuación de regresión estimada y algunas otras informaciones. Lo importante será la interpretación de los resultados que proporcionan estos paquetes de software y no cómo hacer los cálculos para la regresión múltiple.

Un ejemplo: Butler Trucking Company

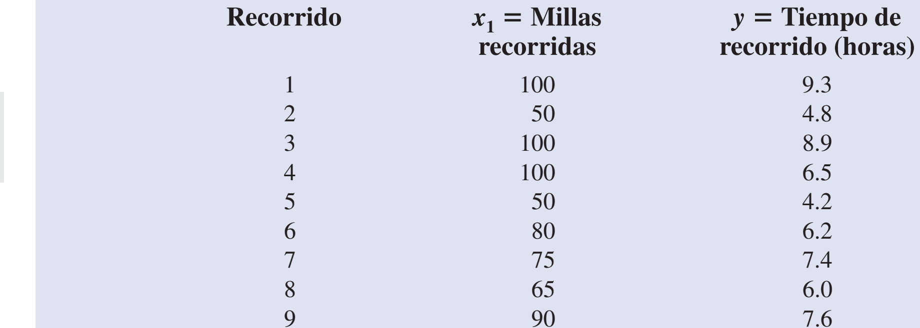
Para ilustrar el análisis de regresión múltiple, se empleará un problema de la empresa Butler Trucking Company, una empresa que se dedica al transporte de objetos y mercancías en el sur de California. La actividad principal de esta empresa es hacer entregas en su área local. Para mejorar el horario de trabajo, los gerentes deseaban estimar el tiempo total de recorrido diario necesario para hacer las entregas.

Al principio, los gerentes creyeron que el tiempo total de recorrido diario estaba estrechamente relacionado con el número de millas recorridas para hacer las entregas. Partiendo de una muestra aleatoria simple de 10 entregas se obtuvieron los datos que se presentan en la tabla 15.1 y en el diagrama de dispersión de la figura 15.2. Después de observar el diagrama de dispersión, los gerentes consideraron que para describir la relación entre tiempo total de recorrido (y_i) y el número de millas recorridas (x_i) podía emplearse el modelo de regresión lineal simple

TABLA 15.1 DATOS PRELIMINARES DE BUTLER TRUCKING



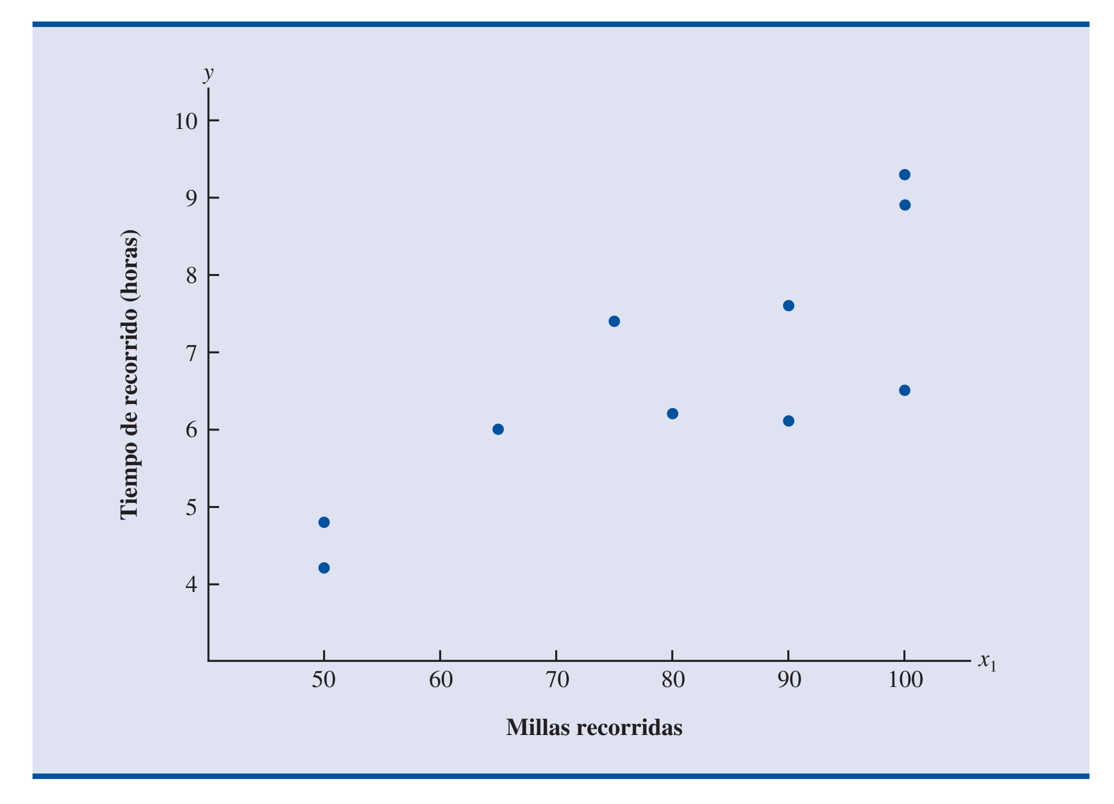
Butler



Recorrido	x_1 = Millas recorridas	y = Tiempo de recorrido (horas)
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1



FIGURA 15.2 DIAGRAMA DE DISPERSIÓN DE LOS DATOS PRELIMINARES DE BUTLER TRUCKING



$y = \beta_0 + \beta_1 x_1 + \epsilon$. Para estimar los parámetros β_0 y β_1, se empleó el método de mínimos cuadrados obteniéndose la ecuación de regresión estimada

$$\hat{y} = b_0 + b_1 x_1 \qquad \textbf{(15.5)}$$

En la figura 15.3 se presentan los resultados obtenidos con Minitab aplicando la regresión lineal simple a los datos de la tabla 15.1. La ecuación de regresión estimada es

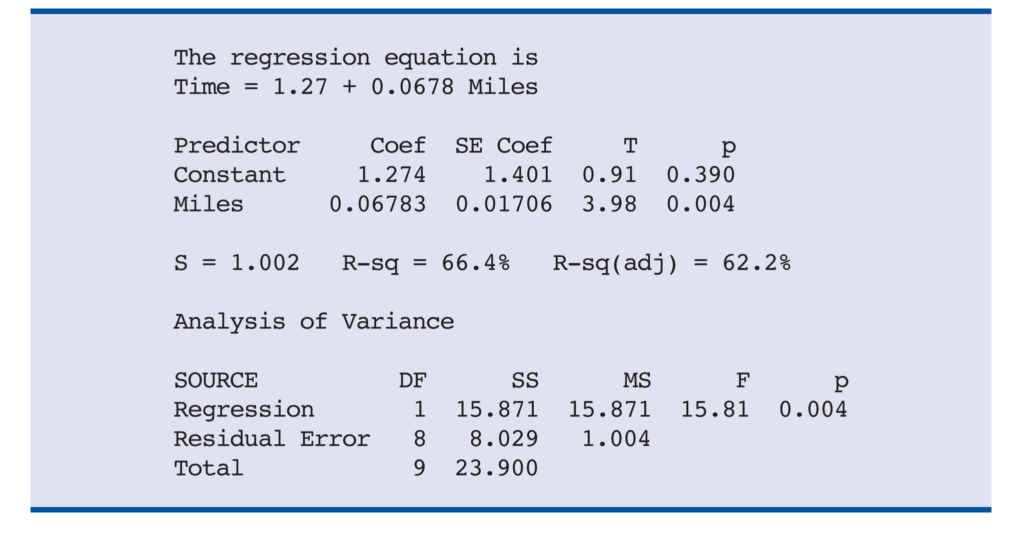
$$\hat{y} = 1.27 + 0.0678 x_1$$

Empleando como nivel de significancia 0.05, el valor-p correspondiente a F de 15.81, es 0.004; esto indica que la relación es significativa, es decir, que se puede rechazar H_0: $\beta_1 = 0$ debido a que el valor-p es menor a $\alpha = 0.05$. Obsérvese que empleando el valor t, 3.98, y su valor-p correspondiente, 0.004, se llega a la misma conclusión. Por lo tanto, se puede concluir que la relación entre el tiempo total de recorrido y el número de millas recorridas es significativa; recorridos de más duración corresponden a cantidades mayores de millas recorridas. Como el coeficiente de correlación (expresado como porcentaje) es R-sq = 66.4%, se ve que 66.4% de la variabilidad en el tiempo de recorrido se puede explicar por el efecto lineal del número de millas recorridas. Este descubrimiento es bastante satisfactorio, sin embargo, los gerentes desean considerar otra variable independiente más para explicar parte de la variabilidad restante de la variable dependiente.

Al tratar de encontrar otra variable independiente los gerentes encontraron que el número de entregas podía contribuir también a la duración total del recorrido. En la tabla 15.2 se presentan los datos de Butler Trucking después de agregar el número de entregas. En la figura 15.4 se presenta el resultado que da Minitab al considerar como variables independientes, tanto el número de millas recorridas (x_1) como el número de entregas (x_2) realizadas. La ecuación de regresión estimada es

$$\hat{y} = -0.869 + 0.0611 x_1 + 0.923 x_2 \qquad \textbf{(15.6)}$$

FIGURA 15.3 RESULTADOS DE MINITAB PARA EL PROBLEMA DE BUTLER TRUCKING CON UNA VARIABLE INDEPENDIENTE



```
The regression equation is
Time = 1.27 + 0.0678 Miles

Predictor      Coef   SE Coef      T      p
Constant      1.274     1.401   0.91  0.390
Miles       0.06783   0.01706   3.98  0.004

S = 1.002   R-sq = 66.4%   R-sq(adj) = 62.2%

Analysis of Variance

SOURCE           DF      SS      MS      F      p
Regression        1  15.871  15.871  15.81  0.004
Residual Error    8   8.029   1.004
Total             9  23.900
```

Los nombres de las variables Miles *(millas) y* Time *(tiempo) que aparecen en los resultados de Minitab fueron ingresados en la hoja de cálculo como encabezados de las columnas correspondientes; por lo tanto,* x_1 = Miles *y* y = Time.

En la sección siguiente se verá el uso del coeficiente de determinación múltiple para medir la bondad de ajuste de esta ecuación de regresión. Pero, antes, se examinarán con más detenimiento los valores $b_1 = 0.0611$ y $b_2 = 0.923$ de la ecuación (15.6).

Nota sobre la interpretación de los coeficientes

En este punto es útil hacer una observación sobre la relación entre la ecuación de regresión estimada en la que la única variable independiente es el número de millas recorridas y la ecuación en la que, como segunda variable independiente, se incluye el número de entregas. El valor de b_1 no es igual en ambos casos. En la regresión lineal simple, b_1 se interpreta como una estimación del cambio en y debido al cambio en una unidad de la variable independiente. En el análisis de regresión múltiple, esta interpretación cambia ligeramente. Es decir, en el análisis de regresión múltiple, cada uno de los coeficientes de regresión se interpreta como sigue: b_i representa la estimación del cambio en y debido a un cambio en una unidad en x_i mientras todas las demás variables independientes permanecen constantes. En el ejemplo de Butler Trucking con dos variables independientes, $b_1 = 0.0611$. Por lo tanto, 0.0611 horas es la estimación del aumen-

TABLA 15.2 DATOS DE BUTLER TRUCKING CON MILLAS RECORRIDAS (x_1) Y CANTIDAD DE ENTREGAS (x_2) COMO VARIABLES INDEPENDIENTES



Recorrido asignado	x_1 = Millas recorridas	x_2 = Cantidad de entregas	y = Tiempo de recorrido (horas)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

FIGURA 15.4 RESULTADOS DE MINITAB PARA EL PROBLEMA DE BUTLER TRUCKING CON DOS VARIABLES INDEPENDIENTES

```
The regression equation is
Time = - 0.869 + 0.0611 Miles + 0.923 Deliveries

Predictor        Coef    SE Coef       T      p
Constant      -0.8687     0.9515   -0.91  0.392
Miles        0.061135   0.009888    6.18  0.000
Deliveries     0.9234     0.2211    4.18  0.004

S = 0.5731   R-sq = 90.4%   R-sq(adj) = 87.6%

Analysis of Variance

SOURCE          DF       SS       MS       F      p
Regression       2   21.601   10.800   32.88  0.000
Residual Error   7    2.299    0.328
Total            9   23.900
```

Los nombres de las variables Miles *(millas),* Deliveries *(entregas) y* Time *(tiempo) que aparecen en los resultados de Minitab fueron ingresados en la hoja de cálculo como encabezados de las columnas; por lo tanto,* x_1 = Miles, x_2 = Deliveries *y* y = Time.

to esperado en el tiempo de recorrido que corresponde al aumento en una milla en la distancia recorrida cuando el número de entregas permanece constante. De manera similar, como $b_2 = 0.923$, una estimación del aumento esperado en el tiempo de recorrido que corresponde al aumento de una entrega permaneciendo constante el número de millas recorridas es 0.923 horas.

Ejercicios

Nota a los estudiantes: Los ejercicios de esta sección y de las secciones siguientes en los que se dan datos están pensados para ser resueltos empleando un paquete de software.

Métodos

1. A continuación se da la ecuación de regresión estimada obtenida a partir de 10 observaciones para un modelo con dos variables independientes.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

   a. Interprete los coeficientes b_1 y b_2 de esta ecuación de regresión estimada.
   b. Estime y para $x_1 = 180$ y $x_2 = 310$.



2. Considérense los datos siguientes que corresponden a la variable dependiente y y a las dos variables independientes x_1 y x_2.



Exer2

x_1	x_2	y
30	12	94
47	10	108
25	17	112
51	16	178
40	5	94
51	19	175
74	7	170

(*continúa*)

x_1	x_2	y
36	12	117
59	13	142
76	16	211

a. Obtenga una ecuación de regresión estimada que relacione y con x_1. Estime y si $x_1 = 45$.
b. Obtenga una ecuación de regresión estimada que relacione y con x_2. Estime y si $x_2 = 15$.
c. Obtenga una ecuación de regresión estimada que relacione y con x_1 y x_2. Estime y si $x_1 = 45$ y $x_2 = 15$.

3. En un análisis de regresión se emplean 30 observaciones y se obtiene la siguiente ecuación de regresión estimada.

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

a. Interprete los coeficientes b_1, b_2, b_3 y b_4 de esta ecuación de regresión.
b. Estime y para $x_1 = 10$, $x_2 = 5$, $x_3 = 1$ y $x_4 = 2$.

Aplicaciones

4. Para una zapatería se obtiene la siguiente ecuación de regresión estimada en la que se relacionan las ventas con la inversión en inventario y los gastos en publicidad.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

donde

$$x_1 = \text{inversión en inventario (en miles de \$)}$$
$$x_2 = \text{gasto en publicidad (en miles de \$)}$$
$$y = \text{ventas (en miles de \$)}$$

a. Estime las ventas si la inversión en inventario es de \$15 000 y el presupuesto para publicidad es de \$10 000.
b. Interprete b_1 y b_2 en esta ecuación de regresión estimada.



5. El dueño de Showtime Movie Theater, Inc., desea estimar el ingreso bruto semanal en función de los gastos en publicidad. A continuación se presentan los datos históricos de 10 semanas.



Showtime

Ingreso semanal bruto (en miles de $)	Publicidad en televisión (en miles de $)	Publicidad en periódicos (en miles de $)
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

a. Obtenga una ecuación de regresión estimada en la que el monto gastado en publicidad en televisión sea la variable independiente.
b. Obtenga una ecuación de regresión estimada en la que los montos gastados en publicidad en televisión y en periódicos sean las variables independientes.
c. ¿Es el coeficiente correspondiente a los gastos de publicidad en televisión de la ecuación de regresión estimada del inciso a) igual al del inciso b)? Interprete este coeficiente en cada caso.

d. ¿Cuál es el ingreso semanal bruto en una semana en la que se gastan $3500 en publicidad en televisión y $1800 en publicidad en periódicos?

6. En el béisbol, el éxito de un equipo se suele considerar en función del desempeño en bateo y en lanzamiento del equipo. Una medida del desempeño en el bateo es la cantidad de cuadrangulares que anota el equipo y una medida del desempeño en lanzamiento es el promedio de carreras ganadas por el equipo que lanza. En general, se cree que los equipos que anotan más cuadrangulares (home run) y tienen un promedio menor de carreras ganadas ganan un mayor porcentaje de juegos. Los datos siguientes pertenecen a 16 equipos que participaron en la temporada de la Liga Mayor de Béisbol de 2003; se da la proporción de juegos ganados, la cantidad de cuadrangulares del equipo (HR, por sus siglas en inglés) y el promedio de carreras ganadas (ERA, por sus siglas en inglés) (www.usatoday.com, 17 de enero de 2004).

archivo en CD

MLB

Equipo	Proporción de ganados	HR	ERA	Equipo	Proporción de ganados	HR	ERA
Arizona	0.519	152	3.857	Milwaukee	0.420	196	5.058
Atlanta	0.623	235	4.106	Montreal	0.512	144	4.027
Chicago	0.543	172	3.842	Nueva York	0.410	124	4.517
Cincinnati	0.426	182	5.127	Philadelphia	0.531	166	4.072
Colorado	0.457	198	5.269	Pittsburgh	0.463	163	4.664
Florida	0.562	157	4.059	San Diego	0.395	128	4.904
Houston	0.537	191	3.880	San Francisco	0.621	180	3.734
Los Ángeles	0.525	124	3.162	St. Louis	0.525	196	4.642

a. Obtenga la ecuación de regresión estimada para predecir la proporción de juegos ganados en función de la cantidad de cuadrangulares.

b. Obtenga la ecuación de regresión estimada para predecir la proporción de juegos ganados en función del promedio de carreras ganadas por los miembros del equipo que lanza.

c. Obtenga la ecuación de regresión estimada para predecir la proporción de juegos ganados en función de la cantidad de cuadrangulares y del promedio de carreras ganadas por los miembros del equipo que lanza.

d. En la temporada de 2003, San Diego ganó sólo el 39.5% de sus juegos, siendo el más bajo de la liga nacional. Para mejorar para el año siguiente, el equipo trató de adquirir nuevos jugadores que hicieran que la cantidad de cuadrangulares aumentara a 180 y que el promedio de carreras ganadas por el equipo que lanza disminuyera a 4.0. Use la ecuación de regresión estimada obtenida en el inciso c) para estimar el porcentaje de juegos que ganaría San Diego si tuviera 180 cuadrangulares y su promedio de carreras ganadas fuera 4.0.

7. Los diseñadores de mochilas usan materiales exóticos como supernailon Derlin, polietileno de alta densidad, aluminio para aviones o espumas termo-moldeadas para hacer que las mochilas sean más confortables y que el peso se distribuya uniformemente eliminándose así los puntos de mayor presión. En los datos siguientes se proporciona capacidad (en pulgadas cúbicas), evaluación del confort, y precio de 10 mochilas probadas por *Outside Magazine.* El confort está medido con una escala del 1 al 5, en la que 1 denota un confort mínimo y 5 un confort excelente. (*Outside Buyer's Guide,* 2001).



Backpack

Fabricante y modelo	Capacidad	Confort	Precio
Camp Trails Paragon II	4330	2	$190
EMS 5500	5500	3	219
Lowe Alpomayo 90+20	5500	4	249
Marmot Muir	4700	3	249
Kelly Bigfoot 5200	5200	4	250
Gregory Whitney	5500	4	340
Osprey 75	4700	4	389
Arc′Teryx Bora 95	5500	5	395
Dana Design Terraplane LTW	5800	5	439
The Works @ Mystery Ranch Jazz	5000	5	525

a. Obtenga la ecuación de regresión estimada que permita predecir el precio de una mochila, dada su capacidad y la evaluación de su confort.
b. Interprete b_1 y b_2.
c. Diga cuál será el precio de una mochila cuya capacidad sea 4500 pulgadas cúbicas y la evaluación de su confort sea 4.

8. En la tabla siguiente se da el rendimiento anual, la evaluación de la seguridad (0 = de alto riesgo, 10 segura) y el coeficiente de gastos anuales de 20 fondos extranjeros (*Mutual Funds,* marzo de 2000).


archivo en CD
ForFunds

	Evaluación de seguridad	Coeficiente de gastos anuales (%)	Rendimiento anual (%)
Accessor Int'l Equity "Adv"	7.1	1.59	49
Aetna "I" International	7.2	1.35	52
Amer Century Int'l Discovery "Inv"	6.8	1.68	89
Columbia International Stock	7.1	1.56	58
Concert Inv "A" Int'l Equity	6.2	2.16	131
Dreyfus Founders Int'l Equity "F"	7.4	1.80	59
Driehaus International Growth	6.5	1.88	99
Excelsior "Inst" Int'l Equity	7.0	0.90	53
Julius Baer International Equity	6.9	1.79	77
Marshall International Stock "Y"	7.2	1.49	54
MassMutual Int'l Equity "S"	7.1	1.05	57
Morgan Grenfell Int'l Sm Cap "Inst"	7.7	1.25	61
New England "A" Int'l Equity	7.0	1.83	88
Pilgrim Int'l Small Cap "A"	7.0	1.94	122
Republic International Equity	7.2	1.09	71
Sit International Growth	6.9	1.50	51
Smith Barney "A" Int'l Equity	7.0	1.28	60
State St Research "S" Int'l Equity	7.1	1.65	50
Strong International Stock	6.5	1.61	93
Vontobel International Equity	7.0	1.50	47

a. Obtenga la ecuación de regresión estimada que relaciona el rendimiento anual con la evaluación de la seguridad y con el coeficiente de gastos anuales.
b. Estime el rendimiento anual de una empresa cuya evaluación de seguridad es 7.5 y el coeficiente de gastos anuales es 2.

9. El ski acuático y el wakeboarding son dos deportes acuáticos muy actuales. Ya sea que se trate de ski acuático, de wakeboarding o de navegación, hallar el modelo que mejor se ajuste a las necesidades, puede no ser una tarea sencilla. La revista *Water Ski* probó 88 lanchas y proporcionó una amplia información como ayuda para los consumidores. A continuación se presenta una parte de los datos que publicaron sobre 20 lanchas de 20 y 22 pies longitud (*Water Ski*, enero/febrero 2006). La manga es el ancho máximo de la lancha (en pulgadas), HP son los caballos de fuerza del motor y velocidad máxima es la velocidad máxima que alcanza la lancha, en millas por hora.



Boats

Fabricante y modelo	Manga	HP	Velocidad máxima
Calabria Cal Air Pro V-2	100	330	45.3
Correct Craft Air Nautique 210	91	330	47.3
Correct Craft Air Nautique SV-211	93	375	46.9
Correct Craft Ski Nautique 206 Limited	91	330	46.7
Gekko GTR 22	96	375	50.1
Gekko GTS 20	83	375	52.2
Malibu Response LXi	93.5	340	47.2
Malibu Sunsettter LXi	98	400	46
Malibu Sunsetter 21 XTi	98	340	44

Fabricante y modelo	Manga	HP	Velocidad máxima
Malibu Sunscape 21 LSV	98	400	47.5
Malibu Wakesetter 21 XTi	98	340	44.9
Malibu Wakesetter VLX	98	400	47.3
Malibu vRide	93.5	340	44.5
Malibu Ride XTi	93.5	320	44.5
Mastercraft ProStar 209	96	350	42.5
Mastercraft X-1	90	310	45.8
Mastercraft X-2	94	310	42.8
Mastercraft X-9	96	350	43.2
MB Sports 190 Plus	92	330	45.3
Svfara SVONE	91	330	47.7

a. Empleando estos datos obtenga la ecuación de regresión estimada que relaciona la velocidad máxima con la manga y los caballos de fuerza de la lancha.
b. La Svfara SV 609 tiene una manga de 85 pulgadas y motor de 330 caballos de fuerza. Utilice la ecuación de regresión estimada obtenida en el inciso a) para estimar la velocidad máxima de la Svfara SV609.

10. La Nacional Basketball Association (NBA) lleva un registro de diversos datos estadísticos de cada equipo. Cuatro de estos datos estadísticos son la proporción de juegos ganados (PCT), la proporción de anotaciones de campo (FG%), la proporción de tiros de tres puntos hechos por el equipo contrario (Opp 3 Pt%) y la cantidad de recuperaciones hechas por el equipo contrario (Opp TO).

Los siguientes datos muestran los valores de estas estadísticas para los 29 equipos de la NBA en una fracción de la temporada 2004 (www.nba.com, enero 3, 2004)



NBA

Equipo	PCT	FG%	Opp 3 Pt%	Opp TO	Equipo	PCT	FG%	Opp 3 Pt%	Opp TO
Atlanta	0.265	0.435	0.346	13.206	Minnesota	0.677	0.473	0.348	13.839
Boston	0.471	0.449	0.369	16.176	Nueva Jersey	0.563	0.435	0.338	17.063
Chicago	0.313	0.417	0.372	15.031	Nueva Orleáns	0.636	0.421	0.330	16.909
Cleveland	0.303	0.438	0.345	12.515	Neuva York	0.412	0.442	0.330	13.588
Dallas	0.581	0.439	0.332	15.000	Orlando	0.242	0.417	0.360	14.242
Denver	0.606	0.431	0.366	17.818	Philadelphia	0.438	0.428	0.364	16.938
Detroit	0.606	0.423	0.262	15.788	Phoenix	0.364	0.438	0.326	16.515
Golden State	0.452	0.445	0.384	14.290	Portland	0.484	0.447	0.367	12.548
Houston	0.548	0.426	0.324	13.161	Sacramento	0.724	0.466	0.327	15.207
Indiana	0.706	0.428	0.317	15.647	San Antonio	0.688	0.429	0.293	15.344
L.A. Clippers	0.464	0.424	0.326	14.357	Seattle	0.533	0.436	0.350	16.767
L.A. Lakers	0.724	0.465	0.323	16.000	Toronto	0.516	0.424	0.314	14.129
Memphis	0.485	0.432	0.358	17.848	Utah	0.531	0.456	0.368	15.469
Miami	0.424	0.410	0.369	14.970	Washington	0.300	0.411	0.341	16.133
Milwaukee	0.500	0.438	0.349	14.750					

a. Obtenga una ecuación de regresión estimada que sirva para predecir la proporción de juegos ganados dada la proporción de anotaciones de campo del equipo.
b. Interprete la pendiente de la ecuación de regresión estimada obtenida en el inciso a).
c. Obtenga una ecuación de regresión estimada que sirva para predecir la proporción de juegos ganados dada la proporción de anotaciones de campo del equipo, la proporción de tiros de tres puntos hechos por el equipo contrario y la proporción de recuperaciones hechas por el equipo contrario.
d. Analice las implicaciones prácticas de la ecuación de regresión estimada obtenida en el inciso c).
e. Estime la proporción de juegos ganados por un equipo para el que los valores de las tres variables independientes son: FG% = 0.45, Opp 3 Pt% = 0.34 y Opp TO = 17.



15.3 Coeficiente de determinación múltiple

En la regresión lineal simple se mostró que la suma de cuadrados se podía dividir o particionar en dos componentes: la suma de cuadrados debida a la regresión y la suma de cuadrados debida al error. Esto también aplica a la suma de cuadrados de la regresión múltiple.

RELACIÓN ENTRE STC, SCR Y SCE

$$\text{STC} = \text{SCR} + \text{SCE} \qquad \textbf{(15.7)}$$

donde

$$\text{STC} = \text{suma total de cuadrados} = \Sigma(y_i - \bar{y})^2$$
$$\text{SCR} = \text{suma de cuadrados debida a la regresión} = \Sigma(\hat{y}_i - \bar{y})^2$$
$$\text{SCE} = \text{suma de cuadrados debida al error} = \Sigma(y_i - \hat{y}_i)^2$$

Debido a lo complejo de los cálculos de estas tres sumas de cuadrados, es necesario emplear un paquete de software para realizarlos. En los resultados de Minitab que se muestran en la figura 15.4, en la parte del análisis de varianza se presentan estos tres valores para el problema de Butler Trucking con dos variables independientes: STC = 23 900, SCR = 21.601 y SCE = 2.299. Cuando se emplea una sola variable independiente (número de millas recorridas) en los resultados de Minitab de la figura 15.3 se observa que STC = 23 900, SCR = 15.871 y SCE = 8.029. El valor de la STC es el mismo en ambos casos debido a que este valor no depende de $\hat{y}$, pero al agregar otra variable (el número de entregas), SCR aumenta y SCE disminuye. Esto tiene como consecuencia que la ecuación de regresión estimada tenga un mejor ajuste a los datos observados.

En el capítulo 14, se empleó el coeficiente de determinación, r^2 = SCR/STC, para medir la bondad de ajuste de la ecuación de regresión estimada. El mismo concepto es válido en la regresión múltiple. El término **coeficiente de determinación múltiple** indica que mide la bondad de ajuste de la ecuación de regresión múltiple estimada. El coeficiente de determinación múltiple, que se denota R^2, se calcula como sigue.

COEFICIENTE DE DETERMINACIÓN MÚLTIPLE

$$R^2 = \frac{\text{SCR}}{\text{STC}} \qquad \textbf{(15.8)}$$

El coeficiente de determinación múltiple puede interpretarse como la proporción de la variabilidad en la variable independiente que es explicada por la ecuación de regresión estimada. Por lo tanto, el producto de este coeficiente por 100, se interpreta como el porcentaje de la variabilidad en y que es explicada por la ecuación de regresión estimada.

Cuando se emplean dos variables independientes en el ejemplo de Butler Truckin, como SCR = 21.601 y STC = 23.900, se tiene

$$R^2 = \frac{21.601}{23.900} = 0.904$$

Por lo tanto, 90.4% de la variabilidad en el tiempo de recorrido y es explicada por la ecuación de regresión estimada en la que las variables independientes son las millas recorridas y el número de entregas. En la figura 15.4 se observa que en el resultado proporcionado por Minitab aparece también el coeficiente de determinación múltiple, que se denota R-sq = 90.4%

Al aumentar el número de variables independientes los errores de predicción se hacen más pequeños, con lo que se reduce la suma de cuadrados debida al error, SCE. Como SCR = STC − SCE, cuando SCE se reduce, SCR aumenta, lo que ocasiona que R^2 = SCR/STC aumente.

En la figura 15.3 el valor de R-sq para la ecuación de regresión estimada con una sola variable, número de millas recorridas (x_1), es 66.4%. Por lo tanto, al agregar el número de entregas como una variable independiente más, el porcentaje de variabilidad en el tiempo de recorrido, explicado por la ecuación de regresión estimada, aumenta de 66.4% a 90.4%. En general, siempre que se agrega una variable independiente al modelo, R^2 aumenta.

Muchos analistas prefieren ajustar R^2 al número de variables independientes para evitar sobreestimar el efecto que tiene agregar una variable independiente sobre la cantidad de la variabilidad explicada por la ecuación de regresión estimada. Siendo n el número de observaciones y p el número de variables independientes, el **coeficiente de determinación ajustado** se calcula como sigue.

Cuando se agrega una variable al modelo, R^2 se vuelve más grande, aun cuando la variable agregada no sea estadísticamente significativa. El coeficiente de determinación múltiple ajustado compensa el número de variables independientes en el modelo.

COEFICIENTE DE DETERMINACIÓN MÚLTIPLE AJUSTADO

$$R_a^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1} \tag{15.9}$$

En el ejemplo de Butler Trucking, como $n = 10$ y $p = 2$, se tiene

$$R_a^2 = 1 - (1 - 0.904)\frac{10 - 1}{10 - 2 - 1} = 0.88$$

Por lo tanto, una vez que el coeficiente de determinación se ha ajustado a dos variables independientes, su valor es 0.88. En los resultados de Minitab de la figura 15.4 este valor se presenta como R-sq(adj) = 87.6%; la diferencia con el valor calculado arriba se debe a que en los cálculos de arriba se empleó un valor redondeado de R^2.

NOTAS Y COMENTARIOS

Si el valor de R^2 es pequeño y el número de variables independientes en el modelo es grande, el coeficiente de determinación ajustado puede ser negativo; en tales casos, Minitab da, como coeficiente de determinación ajustado, cero.

Ejercicios

Métodos

11. En el ejercicio 1 se presentó la siguiente ecuación de regresión estimada basada en 10 0bservaciones.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

Los valores de STC y SCR son 6724.125 y 6216.375, respectivamente.
a. Halle SCE.
b. Calcule R^2.
c. Calcule R_a^2.
d. Analice la bondad de ajuste.



12. En el ejercicio 2, se presentaron 10 observaciones dando los valores de la variable dependiente y y de dos variables independientes x_1 y x_2; con estos datos STC = 15 182.9 y SCR = 14 052.2.
a. Calcule R^2.
b. Calcule R_a^2.
c. ¿Explica la ecuación de regresión estimada una proporción grande de la variabilidad de los datos? Explique.

13. En el ejercicio 3 se presentó la siguiente ecuación de regresión estimada basada en 30 0bservaciones.

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

Los valores de STC y SCR son 1805 y 1760, respectivamente.
a. Calcule R^2.
b. Calcule R_a^2.
c. Analice la bondad de ajuste.

Aplicaciones

14. En el ejercicio 4, se dio la siguiente ecuación de regresión estimada, la cual relacionaba las ventas con la inversión en inventario y los gastos de publicidad.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

Los datos empleados para desarrollar este modelo eran los datos de 10 tiendas; con estos datos STC = 16 000 t SCR y 12 000.
a. Calcule R^2 para la ecuación de regresión estimada.
b. Calcule R_a^2.
c. ¿Parece explicar este modelo una gran cantidad de la variabilidad de los datos? Explique.



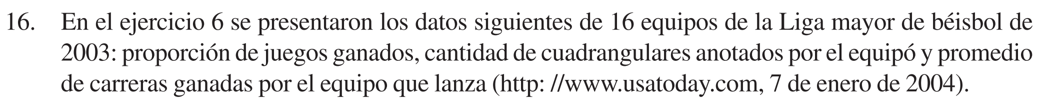
15. En el ejercicio 5, el propietario de Showtime Movie Theater, Inc. empleó el análisis de regresión múltiple para predecir el ingreso bruto (y) en función de la publicidad en televisión (x_1) y de la publicidad en los periódicos (x_2). La ecuación de regresión estimada fue

$$\hat{y} = 83.2 + 2.29x_1 + 1.30x_2$$




La solución obtenida con un paquete de software proporcionó STC = 25.2 y SCR = 23.435.
a. Calcule e interprete R^2 y R_a^2.
b. Cuando la publicidad en televisión era la variable independiente, $R^2 = 0.653$ y $R_a^2 = 0.595$. ¿Prefiere los resultados de la regresión múltiple? Explique.




16. En el ejercicio 6 se presentaron los datos siguientes de 16 equipos de la Liga mayor de béisbol de 2003: proporción de juegos ganados, cantidad de cuadrangulares anotados por el equipó y promedio de carreras ganadas por el equipo que lanza (http: //www.usatoday.com, 7 de enero de 2004).
a. ¿Proporciona un buen ajuste la ecuación de regresión estimada que para predecir la proporción de juegos ganados tiene como única variable independiente la cantidad de cuadrangulares? Explique.
b. Analice la ventaja de usar tanto la cantidad de cuadrangulares como el promedio de carreras ganadas para predecir la proporción de juegos ganados.



17. En el ejercicio 9 se obtuvo una ecuación de regresión estimada que relacionaba la velocidad máxima de una lancha con la manga y los caballos de fuerza de la lancha.
a. Calcule e interprete R^2 y R_a^2.
b. ¿Proporciona esta ecuación de regresión estimada un buen ajuste? Explique.



18. Vuelva al ejercicio 10, en el que se presentaron varios datos estadísticos de 29 equipos de la Nacional Basketball Association en parte de la temporada de 2004 (www.nba.com, 3 de enero de 2004).
a. En el inciso c) del ejercicio 10, se obtuvo una ecuación de regresión estimada que proporcionaba la proporción de juegos ganados dado el porcentaje de anotaciones de campo hechas por el equipo, la proporción de tiros de tres puntos hechas por el equipo contrario y la cantidad de recuperaciones (turnover) hechas por el equipo contrario. ¿Cuáles son los valores de R^2 y R_a^2?
b. ¿Proporciona esta ecuación de regresión estimada un buen ajuste a los datos?



Suposiciones del modelo

En la sección 15.1 se presentó el siguiente modelo de regresión múltiple.

MODELO DE REGRESIÓN MÚLTIPLE

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \quad \textbf{(15.10)}$$

Las suposiciones acerca del término del error ϵ en el modelo de regresión múltiple son análogas a las suposiciones en el modelo de regresión lineal simple.

SUPOSICIONES SOBRE EL TÉRMINO DEL ERROR ϵ EN EL MODELO DE REGRESIÓN MÚLTIPLE $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$

1. El término del error ϵ es una variable aleatoria cuya media o valor esperado es cero, es decir, $E(\epsilon) = 0$.
   Consecuencia: Para valores dados de $x_1, x_2, \ldots, x_p$, el valor esperado o valor promedio de y está dado por

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \quad \textbf{(15.11)}$$

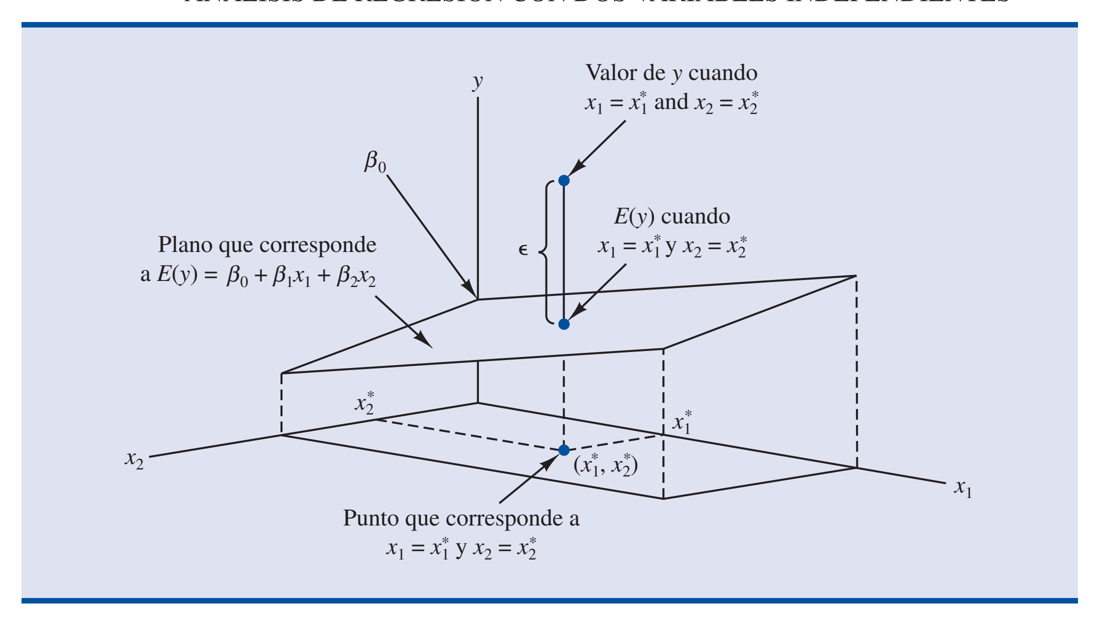
   La ecuación (15.11) es la ecuación de regresión múltiple presentada en la sección 15.1. En esta ecuación, $E(y)$ representa el promedio de todos los valores que puede tomar y para valores dados de $x_1, x_2, \ldots, x_p$.
2. La varianza de ϵ se denota σ^2 y es la misma para todos los valores de las variables independientes $x_1, x_2, \ldots, x_p$.
   Consecuencia: La varianza de y respecto a la línea de regresión es σ^2 y es la misma para todos los valores de $x_1, x_2, \ldots, x_p$.
3. Los valores de ϵ son independientes.
   Consecuencia: El valor de ϵ para un determinado conjunto de valores de las variables independientes no está relacionado con el valor de ϵ de ningún otro conjunto de valores.
4. El término del error ϵ es una variable aleatoria distribuida normalmente y que refleja la desviación entre el valor de y y el valor esperado de y dado por $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.
   Consecuencia: Como $\beta_0, \beta_1, \ldots, \beta_p$ son constantes para los valores dados de $x_1, x_2, \ldots, x_p$, la variable dependiente y es también una variable aleatoria distribuida normalmente.

Para entender mejor la forma de la relación dada por la ecuación 15.11, considérese la siguiente ecuación de regresión múltiple con dos variables independientes.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

La gráfica de esta ecuación es un plano en el espacio tridimensional. La figura 15.5 es un ejemplo de gráfica de este tipo. Obsérvese que, como se indica, el valor de ϵ es la diferencia entre el verdadero valor de y y el valor esperado de y, $E(y)$, cuando $x_1 = x_1^*$ y $x_2 = x_2^*$.

FIGURA 15.5 GRÁFICA DE LA ECUACIÓN DE REGRESIÓN EMPLEADA EN EL ANÁLISIS DE REGRESIÓN CON DOS VARIABLES INDEPENDIENTES



En el análisis de regresión, suele emplearse del término *variable de respuesta* en lugar del término *variable dependiente.* Además, como la ecuación de regresión múltiple genera un plano o superficie, a su gráfica se le llama superficie de respuesta.



Prueba de significancia

En esta sección se muestra cómo realizar una prueba de significancia para una relación de regresión múltiple. Las pruebas de significancia que se usaron en la regresión lineal simple fueron la prueba t y la prueba F. En la regresión lineal simple, estas dos pruebas llevan a la misma conclusión; es decir, si se rechaza la hipótesis nula, se concluye que $b_1 \neq 0$. En la regresión múltiple, la prueba t y la prueba F tienen propósitos diferentes.

1. La prueba F se usa para determinar si existe una relación de significancia entre la variable dependiente y el conjunto de todas las variables independientes; a esta prueba F se le llama prueba de *significancia global*.
2. Si la prueba F indica que hay significancia global, se usa la prueba t para ver si cada una de las variables individuales es significativa. Para cada una de las variables independientes del modelo se realiza una prueba t. A cada una de estas pruebas t se le conoce como prueba de *significancia individual.*

A continuación se explican la prueba F y la prueba t y se aplican al ejemplo de Butler Trucking.

Prueba *F*

El modelo de regresión múltiple que se definió en la sección 15.4 es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

La hipótesis de la prueba F comprende los parámetros del modelo de regresión múltiple.

H_0: $\beta_1 = \beta_2 = \quad = \beta_p = 0$

H_a: Uno o más de los parámetros es distinto de cero

Cuando se rechaza H_0, la prueba proporciona evidencia estadística suficiente para concluir que uno o más de los parámetros no es igual a cero y que la relación global entre y y el conjunto de variables independientes $x_1, x_2, \ldots, x_p$ es significativa. En cambio, si no se puede rechazar H_0, no se tiene evidencia suficiente para concluir que exista una relación significativa.

Antes de describir los pasos de la prueba F, es necesario revisar el concepto de *cuadrado medio*. Un cuadrado medio es una suma de cuadrados dividida entre sus correspondientes grados de libertad. En el caso de la regresión múltiple, la suma de cuadrados del total tiene $n - 1$ grados de libertad, la suma de cuadrados debida a la regresión (SCR) tiene p grados de libertad y la suma de cuadrados debida al error tiene $n - p - 1$ grados de libertad. Por tanto, el cuadrado medio debido a la regresión (CMR) es SCR/p y el cuadrado medio debido al error (CME) es SCE/$(n - p - 1)$.

$$\text{CMR} = \frac{\text{SCR}}{p} \tag{15.12}$$

y

$$\text{CME} = \frac{\text{SCE}}{n - p - 1} \tag{15.13}$$

Como se vio en el capítulo 14, CME proporciona una estimación insesgada de σ^2, la varianza del término del error ϵ. Si H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ es verdadera, CMR también proporciona un estimador insesgado de σ^2 y el valor de CMR/CME será cercano a 1. Pero, si H_0 es falsa, el CMR sobreestima σ^2 y el valor de CMR/CME será mayor. Para determinar qué tan grande debe ser CMR/CME para que se rechace H_0, se hace uso del hecho de que si H_0 es verdadera y las suposiciones acerca del modelo de regresión múltiple son válidas, la distribución muestral de CMR/CME es una distribución F con p grados de libertad en el numerador y $n - p - 1$ en el denominador. A continuación se presenta un resumen de la prueba F de significancia para la regresión múltiple.

PRUEBA F DE SIGNIFICANCIA GLOBAL

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
H_a: Uno o más de los parámetros no son iguales a cero.

ESTADÍSTICO DE PRUEBA

$$F = \frac{\text{CMR}}{\text{CME}} \tag{15.14}$$

REGLA DE RECHAZO

Valor aproximado p:	Rechazar H_0 si valor $p \leq \alpha$
Valor crítico aproximado:	Rechazar H_0 si $F \geq F_\alpha$

donde F_α pertenece a la distribución F con p grados de libertad en el numerador y $n - p - 1$ grados de libertad en el denominador.

A continuación se presenta la aplicación de la prueba F al problema de regresión múltiple de la empresa Butler Trucking. Como se tienen dos variables independientes, las hipótesis se expresan como sigue.

H_0: $\beta_1 = \beta_2 = 0$
H_a: β_1 o β_2 no es igual a cero

FIGURA 15.6 RESULTADOS DE MINITAB PARA EL EJEMPLO DE BUTLER TRUCKING CON DOS VARIABLES INDEPENDIENTES, MILLAS RECORRIDAS (x_1) Y NÚMERO DE ENTREGAS (x_2)

```
The regression equation is
Time = - 0.869 + 0.0611 Miles + 0.923 Deliveries

Predictor        Coef   SE Coef       T      p
Constant      -0.8687    0.9515   -0.91  0.392
Miles        0.061135  0.009888    6.18  0.000
Deliveries     0.9234    0.2211    4.18  0.004

S = 0.5731   R-sq = 90.4%   R-sq(adj) = 87.6%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       2  21.601  10.800  32.88  0.000
Residual Error   7   2.299   0.328
Total            9  23.900
```

En la figura 15.6 se presentan los resultados que da Minitab para el modelo de regresión múltiple con dos variables independientes, millas recorridas (x_1) y número de entregas (x_2). En la parte de los resultados que corresponde al análisis de varianza, se ve que CMR $=$ 10.8 y CME $=$ 0.328. Empleando la ecuación (15.14) se obtiene el valor del estadístico de prueba.

$$F = \frac{10.8}{0.328} = 32.9$$

Obsérvese que el valor de F en los resultados de Minitab es $F = 32.88$; este valor difiere del calculado aquí debido a que aquí en los cálculos se emplearon los valores redondeados de CMR y CME. Usando $\alpha = 0.01$, el valor-$p = 0.000$ que aparece en la última columna de la tabla del análisis de varianza (figura 15.6) indica que se puede rechazar H_0: $\beta_1 = \beta_2 = 0$ debido a que el valor-p es menor a 0.01. De manera alternativa, en la tabla 4 del apéndice B se observa que con dos grados de libertad en el numerador y siete grados de libertad en el denominador, $F_{0.01} = 9.55$. Como $32.9 > 9.55$, se rechaza H_0: $\beta_1 = \beta_2 = 0$ y se concluye que existe una relación significativa entre el tiempo de recorrido y y las dos variables independientes, millas recorridas y número de entregas.

Como ya se indicó, el error cuadrado medio proporciona un estimador insesgado de σ^2, la varianza del término del error ϵ. Observando la figura 15.6 se ve que la estimación de σ^2 es CME = 0.328. La raíz cuadrada del CME es la estimación de la desviación estándar del término del error. Como se definió en la sección 14.5, esta desviación estándar es el error estándar de estimación, que se denota s. Por tanto, se tiene que $s = \sqrt{\text{CME}} = \sqrt{0.328} = 0.573$. Observe que este valor del error estándar de estimación aparece en los resultados de Minitab que se muestran en la figura 15.6.

La tabla 15.3 es la tabla general para el análisis de varianza (ANOVA) que proporciona los resultados de la prueba F para un modelo de regresión múltiple. El valor del estadístico de prueba F aparece en la última columna y debe compararse con F_α con p grados de libertad en el numerador y $n - p - 1$ grados de libertad en el denominador, para obtener la conclusión de la prueba de hipótesis. Revisando los resultados de Minitab para el ejemplo de Butler Trucker Company de la figura 15.6 se ve que la tabla del análisis de varianza contiene esta información. Además, Minitab proporciona también el correspondiente valor-p para el estadístico de prueba F.

TABLA 15.3 TABLA ANOVA PARA EL MODELO DE REGRESIÓN MÚLTIPLE CON p VARIABLES INDEPENDIENTES

Fuente	Suma de cuadrados	Grados de libertad	Cuadrado medio	F
Regresión	SCR	p	$CMR = \frac{SCR}{p}$	$F = \frac{CMR}{CME}$
Error	SCE	$n - p - 1$	$CME = \frac{SCE}{n - p - 1}$	
Total	STC	$n - 1$		

Prueba t

Si la prueba F indica que la relación de regresión múltiple es significativa, se puede realizar una prueba t para determinar la significancia de cada uno de los parámetros. A continuación se presenta la prueba t de significancia para cada uno de los parámetros.

PRUEBA t DE SIGNIFICANCIA PARA CADA UNO DE LOS PARÁMETROS

Para cualquier parámetro β_i

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

ESTADÍSTICO DE PRUEBA

$$t = \frac{b_i}{s_{b_i}} \tag{15.15}$$

REGLA DE RECHAZO

Método del valor-p: Rechazar H_0 si valor-$p \leq \alpha$
Método del valor crítico: Rechazar H_0 si $t \leq -t_{\alpha/2}$ o si $t \geq t_{\alpha/2}$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - p - 1$ grados de libertad.

En el estadístico de prueba, s_{b_i} es la estimación de la desviación estándar de b_i. El paquete de software proporciona el valor de s_{b_i}.

A continuación se realiza la prueba t utilizando el problema de regresión de Butler Trucking. Consúltese la sección de la figura 15.6 en la que se dan los resultados de Minitab para el cálculo del cociente t. Los valores de b_1, s_{b_1}, b_2 y s_{b_2}, son los siguientes

$$b_1 = 0.061135 \qquad s_{b_1} = 0.009888$$
$$b_2 = 0.9234 \qquad s_{b_2} = 0.2211$$

Usando la ecuación (15.15), se obtienen los estadísticos de prueba para las hipótesis en que intervienen β_1 y β_2

$$t = 0.061135/0.009888 = 6.18$$
$$t = 0.9234/0.2211 = 4.18$$

Obsérvese que los valores de estos dos cocientes t y sus correspondientes valores-p aparecen en los resultados de Minitab de la figura 15.6. Usando, $\alpha = 0.01$, los valores-p 0.000 y 0.004 que aparecen en los resultados de Minitab indican que se pueden rechazar H_0: $\beta_1 = 0$ y H_0: $\beta_2 = 0$. Por lo tanto, ambos parámetros son estadísticamente significativos. También, en la tabla 2 del apéndice B se encuentra que para $n - p - 1 = 10 - 2 - 1 = 7$ grados de libertad, $t_{0.005} = 3.499$. Como $6.18 > 3.499$, se rechaza H_0: $\beta_1 = 0$. De manera similar, como $4.18 > 3.499$, se rechaza H_0: $\beta_2 = 0$.

Multicolinealidad

En el análisis de regresión el término *variable independiente* se usa para referirse a cualquier variable que se usa para predecir o explicar el valor de la variable dependiente. Sin embargo, este término no significa que estas variables independientes sean independientes entre ellas, en sentido estadístico. Al contrario, en un problema de regresión múltiple, la mayoría de las variables independientes están, en cierto grado, correlacionadas unas con otras. En el ejemplo de Butler Trucking con dos variables independientes x_1 (millas recorridas) y x_2 (número de entregas), las millas recorridas pueden tratarse como la variable dependiente y el número de entregas como la variable independiente para determinar si estas dos variables están relacionadas entre sí. Después puede calcularse el coeficiente de correlación muestral $r_{x_1x_2}$ para determinar la magnitud de la relación entre estas dos variables. Al hacer esto se obtiene $r_{x_1x_2} = 0.16$. Por lo tanto, se encuentra que existe cierto grado de relación lineal entre estas dos variables independientes. En el análisis de regresión múltiple **multicolinealidad** se refiere a la correlación entre las variables independientes.

Para tener una mejor visión de los problemas potenciales de la multicolinealidad, se considerará una modificación al ejemplo de Butler Trucking. En lugar de que x_2 sea el número de entregas, sea x_2 el número de galones de gasolina consumidos. Es claro que x_1 (las millas recorridas) y x_2 están relacionadas, es decir, se sabe que el número de galones de gasolina consumidos depende del número de millas recorridas. Por lo tanto, se concluirá que x_1 y x_2 son variables independientes fuertemente correlacionadas.

Supóngase que se obtiene la ecuación $\hat{y} = b_0 + b_1x_1 + b_2x_2$ y que se encuentra que la prueba F indica que esta relación es significativa. Después supóngase que se realiza la prueba t para β_1 para determinar si $\beta_1 \neq 0$ y no puede rechazarse H_0: $\beta_1 = 0$. ¿Significa esto que el tiempo de recorrido no esté relacionado con las millas recorridas? No necesariamente. Lo que probablemente significa es que estando x_2 en el modelo, x_1 no tiene una contribución significativa en la determinación del valor de y. En el presente ejemplo, esta interpretación parece razonable; conociendo la cantidad de gasolina consumida, no se gana mucha más información para la predicción de y al conocer el número de millas recorridas. De manera similar, una prueba t puede llevar a la conclusión de que $\beta_2 = 0$ con base en que, cuando está x_1 en el modelo, al saber la cantidad de gasolina consumida, no se gana mucho.

Una regla práctica que previene de potenciales problemas de multicolinealidad es que el coeficiente de correlación muestral sea mayor a + 0.7 o menor a − 0.7

Resumiendo, en las pruebas t para la significancia de cada uno de los parámetros, la dificultad ocasionada por la multicolinealidad hace posible concluir que ninguno de los parámetros es significativamente distinto de cero cuando la prueba F sobre la ecuación de regresión múltiple general indica que hay una relación significante. Este problema se evita cuando existe poca correlación entre las variables independientes.

Para determinar si la multicolinealidad es lo suficientemente alta para ocasionar problemas se han desarrollado diversas pruebas. De acuerdo con la prueba de la regla práctica, la multicolinealidad es un problema potencial si el valor absoluto del coeficiente de correlación muestral es mayor a 0.7 para cualquier par de variables independientes. Otros tipos de pruebas son más avanzadas y quedan fuera del alcance de este libro.

Cuando las variables independientes están fuertemente correlacionadas, es imposible determinar el efecto, por separado, de cada una de las variables independientes sobre la variable dependiente.

Siempre que sea posible, debe evitarse incluir variables independientes que estén fuertemente correlacionadas. Sin embargo, en la práctica, la estricta adherencia a esta conducta no suele ser posible. Cuando las personas que toman las decisiones tienen razones para creer que existe una multicolinealidad importante, se darán cuenta de que es difícil separar los efectos de cada una de las variables independientes sobre la variable dependiente

NOTAS Y COMENTARIOS

Por lo general, la multicolinealidad no afecta la manera en que se realiza el análisis de regresión o en que se interpretan los resultados de un estudio. Pero, si la multicolinealidad es severa se pueden tener dificultades al interpretar los resultados de las pruebas t acerca de cada uno de los parámetros. Además del tipo de problemas ilustrados en esta sección, se ha demostrado que los casos severos de multicolinealidad dan como resultado estimaciones por mínimos cuadrados con signo erróneo. Esto es, en estudios simulados en los que los investigadores crearon el modelo de regresión subyacente y después emplearon el método de mínimos cuadrados para obtener estimaciones de β_0, β_1, β, etc., se ha demostrado que en condiciones de fuerte multicolinealidad, las estimaciones obtenidas por mínimos cuadrados pueden tener signo opuesto al del parámetro que se estima. Por ejemplo, β_2 puede ser en realidad $+10$ y su estimación b_2 puede resultar ser -2. Por lo tanto, si existe una fuerte multicolinealidad podrá tenerse poca confianza en los coeficientes.

Ejercicios

Métodos



19. En el ejercicio 1 se presentó la siguiente ecuación de regresión estimada basada en 10 observaciones.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

Aquí STC $= 6724.125$, SCR $= 6216.375$, $s_{b_1} = 0.0813$ y $s_{b_2} = 0.0567$.

a. Calcule CMR y CME
b. Calcule F y realice la prueba F adecuada. Use $\alpha = 0.05$.
c. Realice una prueba t para la significancia de b_1. Use $\alpha = 0.05$.
d. Realice una prueba t para la significancia de b_2. Use $\alpha = 0.05$.

20. Consulte los datos presentados en el ejercicio 2. La ecuación de regresión estimada de estos datos es

$$\hat{y} = -18.4 + 2.01x_1 + 4.74x_2$$

Aquí STC = 15 182, SCR = 14 052. 2, $s_{b_1} = 0.2471$.

a. Realice una prueba para ver si hay una relación significativa entre x_1, x_2 y y. Use $\alpha = 0.05$.
b. ¿Es significativo β_1? Use $\alpha = 0.05$.
c. ¿Es significativo β_2? Use $\alpha = 0.05$.

21. Se obtuvo la ecuación de regresión estimada siguiente para un modelo con dos variables independientes.

$$\hat{y} = 40.7 + 8.63x_1 + 2.71x_2$$

Después de eliminar x_2 del modelo, se empleó el método de mínimos cuadrados para obtener una ecuación de regresión estimada con una sola variable independiente, x_1.

$$\hat{y} = 42.0 + 9.01x_1$$

a. Dé la interpretación del coeficiente de x_1 en ambos modelos.
b. ¿Podría explicar la multicolinealidad por qué el coeficiente de x_1 es diferente en los dos modelos? Si es así, ¿cómo?

Aplicaciones

22. En el ejercicio 4 se dio la siguiente ecuación de regresión estimada que relacionaba las ventas con la inversión en inventario y los gastos de publicidad.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

Los datos empleados para obtener el modelo provinieron de un estudio realizado a 10 tiendas; para estos datos STC = 16 000 y SCR = 12 000.

a. Calcule SCE, CME y CMR.
b. Use la prueba F y 0.05 como nivel de significancia para determinar si existe una relación entre las variables.

Autoexamen

23. Véase el ejercicio 5.

a. Use $\alpha = 0.01$ para probar las hipótesis

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_a: \beta_1 \text{ o } \beta_2 \text{ no son iguales a cero}$$

En el modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, donde

$$x_1 = \text{publicidad en televisión (en miles de dólares)}$$
$$x_2 = \text{publicidad en periódicos (en miles de dólares)}$$

b. Use $\alpha = 0.05$ para probar la significancia de β_1 ¿Debe x_1 ser eliminada del modelo?
c. Use $\alpha = 0.05$ para probar la significancia de β_2 ¿Debe x_2 ser eliminada del modelo?

24. Véanse los datos del ejercicio 6. Emplee la cantidad de cuadrangulares del equipo y el promedio de carreras ganadas por el equipo que lanza para predecir la proporción de juegos ganados.

a. Use la prueba F para determinar la significancia global de la relación. ¿Cuál es la conclusión empleando 0.05 como nivel de significancia?
b. Use la prueba t para determinar la significancia de cada una de las variables independientes. ¿Cuál es la conclusión empleando 0.05 como nivel de significancia?



25. *Borron's* realiza revisiones anuales de los corredores de bolsa en línea, en la que se incluyen tanto corredores a los que se les puede contactar vía un explorador de Internet, así como corredores que tienen acceso directo y que ponen al cliente en contacto directo con el servidor de una red de corredores de bolsa. La oferta y el desempeño de cada corredor se evalúa en seis áreas, empleando para cada área una escala de 0 a 5. Los resultados se ponderan para obtener una evaluación general y a cada corredor se le asigna una evaluación final que va de cero a cinco estrellas. Tres de las áreas evaluadas son ejecución de la operación, facilidad de uso y gama de ofertas. Un 5 en ejecución de la operación significa que la llegada del pedido y el proceso de ejecución fluyó con facilidad de un paso a otro. En facilidad de uso, 5 significa que el sitio es de fácil uso y que se puede ajustar para ver lo que le interesa al usuario ver. Un 5 en gama de ofertas significa que todas las transacciones pueden realizarse en línea. En los datos siguientes se presentan las puntuaciones obtenidas en ejecución de la operación, facilidad de uso y gama de ofertas y el número de estrellas obtenidas por los integrantes de una muestra de 10 corredores de bolsa (*Barron's*, 10 de marzo de 2003).

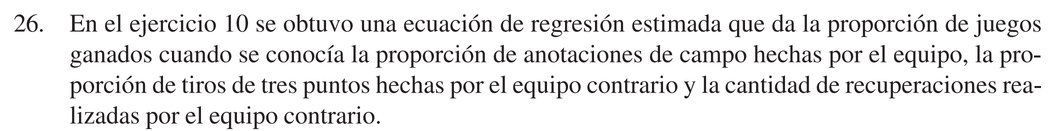


Corredor	Ejecución de la operación	Uso	Gama	Estrellas
Wall St. Access	3.7	4.5	4.8	4.0
E*TRADE (Power)	3.4	3.0	4.2	3.5
E*TRADE (Standard)	2.5	4.0	4.0	3.5
Preferred Trade	4.8	3.7	3.4	3.5
my Track	4.0	3.5	3.2	3.5
TD Waterhouse	3.0	3.0	4.6	3.5
Brown & Co.	2.7	2.5	3.3	3.0
Brokerage America	1.7	3.5	3.1	3.0
Merrill Lynch Direct	2.2	2.7	3.0	2.5
Strong Funds	1.4	3.6	2.5	2.0

a. Determine la ecuación de regresión estimada que se puede usar para predecir el número de estrellas dadas las evaluaciones a ejecución, facilidad de uso y gama de ofertas.
b. Emplee la prueba F para determinar la significancia global de la relación. Empleando como nivel de significancia 0.95, ¿cuál es la conclusión?
c. Emplee la prueba t para determinar la significancia de cada una de las variables independientes. Empleando como nivel de significancia 0.05, ¿cuál es la conclusión?
d. Elimine cualquiera de las variables independientes que no sea significativa para la ecuación de regresión estimada. ¿Cuál es la ecuación de regresión estimada que recomienda? Compare R^2 con el valor de R^2 para el inciso a). Analice las diferencias.



NBA



26. En el ejercicio 10 se obtuvo una ecuación de regresión estimada que da la proporción de juegos ganados cuando se conocía la proporción de anotaciones de campo hechas por el equipo, la proporción de tiros de tres puntos hechas por el equipo contrario y la cantidad de recuperaciones realizadas por el equipo contrario.
a. Emplee la prueba F para determinar la significancia global de la relación. Empleando como nivel de significancia 0.05, ¿cuál es la conclusión?
b. Emplee la prueba t para determinar la significancia de cada una de las variables independientes. Empleando como nivel de significancia 0.05, ¿cuál es la conclusión?



Uso de la ecuación de regresión estimada para estimaciones y predicciones

Los procedimientos empleados en la regresión múltiple para estimar el valor medio de y y para predecir el valor de un solo valor de y son similares a los empleados en el análisis de regresión para una sola variable independiente. Recuérdese, primero, que en el capítulo 14 se mostró que la estimación puntual del valor esperado de y para un valor dado de x y la estimación puntual de un solo valor de y es la misma. En ambos casos se usó como estimación puntual $\hat{y} = b_0 + b_1x$.

En la regresión múltiple se emplea el mismo procedimiento, es decir, los valores dados de x_1, $x_2, \ldots, x_p$ se sustituyen en la ecuación de regresión y como estimación puntual se emplea el correspondiente valor de $\hat{y}$. Supóngase que en el ejemplo de Butler Truking se desea usar la ecuación de regresión estimada con x_1 (millas recorridas) y x_2 (número de entregas) para obtener dos estimaciones por intervalo:

1. Un *intervalo de confianza* para la media del tiempo de recorrido de todos los camiones que recorren 100 millas y hacen dos entregas.
2. Un *intervalo de predicción* para el tiempo de recorrido de un determinado camión que recorre 100 millas y hace dos entregas.

Utilizando la ecuación de regresión estimada $\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$ con $x_1 = 100$ y $x_2 = 2$, se obtiene

$$\hat{y} = -0.869 + 0.0611(100) + 0.923(2) = 7.09$$

Por lo tanto, en ambos casos, la estimación puntual del tiempo de recorrido es aproximadamente 7 horas.

Para obtener las estimaciones por intervalo para el valor medio de y y para un solo valor de y se emplean procedimientos similares a los empleados en el análisis de regresión con una sola variable. Las fórmulas que se necesitan quedan fuera del alcance de este libro, sin embargo, los paquetes de software para el análisis de regresión múltiple suelen proporcionar intervalos de confianza, una vez que los valores $x_1, x_2, \ldots, x_p$ hayan sido especificados por el usuario. En la tabla 15.4 se presentan los intervalos de confianza y de predicción para algunos valores de x_1 y x_2 del ejemplo de Butler Trucking; estos valores se obtuvieron empleando Minitab. Obsérvese que las estimaciones por intervalo para un solo valor de y proporcionan valores más amplios que las estimaciones por intervalo para el valor esperado de y. Esta diferencia refleja simplemente el hecho de que dados los valores x_1 y x_2, se puede estimar con mayor precisión el tiempo medio de recorrido de todos los camiones, que predecir el tiempo medio de recorrido de un determinado camión.

TABLA 15.4 INTERVALOS DE 95% DE CONFIANZA Y DE PREDICCIÓN PARA EL EJEMPLO DE BUTLER TRUCKING

Valor de x_1	Valor de x_2	Intervalo de confianza		Intervalo de predicción	
		Límite inferior	Límite superior	Límite inferior	Límite superior
50	2	3.146	4.924	2.414	5.656
50	3	4.127	5.789	3.368	6.548
50	4	4.815	6.948	4.157	7.607
100	2	6.258	7.926	5.500	8.683
100	3	7.385	8.645	6.520	9.510
100	4	8.135	9.742	7.362	10.515

Ejercicios

Métodos

27. En el ejercicio 1 se presentó la siguiente ecuación de regresión estimada basada en 10 observaciones.

$$\hat{y} = 29.1270 + 0.5906x_1 + 0.4980x_2$$

a. Obtenga una estimación puntual del valor medio de y para $x_1 = 180$ y $x_2 = 310$.
b. Obtenga una estimación puntual para un solo valor de y siendo $x_1 = 180$ y $x_2 = 310$.



28. Véanse los datos del ejercicio 2. La ecuación de regresión estimada para estos datos es

$$\hat{y} = -18.4 + 2.01x_1 + 4.74x_2$$

a. Obtenga un intervalo de 95% de confianza para el valor medio de y cuando $x_1 = 45$ y $x_2 = 15$.
b. Obtenga un intervalo de predicción de 95% para el valor de y cuando $x_1 = 45$ y $x_2 = 15$.

Aplicaciones



29. En el ejercicio 5, el propietario de Showtime Movie Theater, Inc. empleó el análisis de regresión múltiple para predecir el ingreso bruto (y) en función de la publicidad en televisión (x_1) y de la publicidad en periódicos (x_2). La ecuación de regresión estimada fue

$$\hat{y} = 83.2 + 2.29x_1 + 1.30x_2$$

a. ¿Cuál será el ingreso bruto esperado en una semana en la que se gastan $3500 en publicidad en televisión ($x_1 = 3.5$) y $1800 en publicidad en periódicos ($x_2 = 1.8$)?
b. Dé un intervalo de 95% de confianza para el ingreso medio de todas las semanas en las que los gastos sean los indicados en el inciso a).
c. Dé un intervalo de predicción de 95% para la media del ingreso de una semana en las que los gastos sean los indicados en el inciso a).



30. En el ejercicio 9 se obtuvo una ecuación de regresión estimada que relacionaba la máxima velocidad de una lancha con su manga y sus caballos de fuerza.
a. Dé un intervalo de 95% de confianza para la media de la velocidad máxima de una lancha cuya manga sea 85 y cuyo motor tenga 330 caballos de fuerza.

b. La manga de la Svfara SV 609 es de 85 pulgadas y su motor tiene 330 caballos de fuerza. Dé un intervalo de 95% de confianza para la media de la velocidad máxima de la Svfara 609.

31. La sección "Guía para el usuario" del sitio en la Red de la revista *Car and Driver* proporciona información sobre pruebas viales (road test) de automóviles, camiones, SUV (acrónimo en inglés de Sport Utility Vehicle) y vans. Abajo se presentan las puntuaciones generales para calidad general, modelo de vehículo, frenado, manejo, economía de combustible, confort interior, aceleración, confiabilidad, ajuste y terminado, transmisión dadas a diversos vehículos empleando una escala del 1 (lo peor) a 10 (lo mejor). Aquí se presenta una parte de los datos de 14 automóviles Deportivos/GT (www.caranddriver.com, 7 de enero de 2004).



Deportivos/GT	General	Manejo	Confiabilidad	Ajuste y terminado
Acura 3.2CL	7.80	7.83	8.17	7.67
Acura RSX	9.02	9.46	9.35	8.97
Audi TT	9.00	9.58	8.74	9.38
BMW 3-Series/M3	8.39	9.52	8.39	8.55
Chevrolet Corvette	8.82	9.64	8.54	7.87
Ford Mustang	8.34	8.85	8.70	7.34
Honda Civic Si	8.92	9.31	9.50	7.93
Infiniti G35	8.70	9.34	8.96	8.07
Mazda RX-8	8.58	9.79	8.96	8.12
Mini Cooper	8.76	10.00	8.69	8.33
Mitsubishi Eclipse	8.17	8.95	8.25	7.36
Nissan 350Z	8.07	9.35	7.56	8.21
Porsche 911	9.55	9.91	8.86	9.55
Toyota Celica	8.77	9.29	9.04	7.97

a. Dé una ecuación de regresión estimada usando manejo, confiabilidad, y ajuste y terminado para predecir la calidad general.
b. Otro de los automóviles deportivos/GT evaluados por *Car and Driver* es el Honda Accord. Las evaluaciones de manejo, confiabilidad, y ajuste y terminado dadas a este automóvil fueron 8.28, 9.06 y 8.07, respectivamente. Estime la evaluación general dada a este automóvil.
c. Dé un intervalo de 95% de confianza para la calidad general de todos los automóviles deportivos y GT con las características enumeradas en el inciso a).
d. Dé un intervalo de predicción de 95% para la calidad general del Honda Accord descrito en el inciso b).
e. La evaluación general dada por *Car and Driver* para el Honda Accord fue 8.65. Compare esta evaluación con las estimaciones obtenidas en los incisos b) y d).



Variables cualitativas independientes

Las variables independientes pueden ser cualitativas o cuantitativas.

En los ejemplos considerados hasta ahora, las variables han sido variables independientes cuantitativas como, por ejemplo, población de estudiantes, distancia recorrida y número de entregas. Sin embargo, en muchas situaciones, se tiene que trabajar con **variables independientes cualitativas** como género (masculino o femenino), modo de pago (efectivo, tarjeta de crédito, cheque), etc. En esta sección, el objetivo es mostrar cómo se emplean las variables cualitativas en el análisis de regresión. Para ilustrar el uso e interpretación de las variables independientes cualitativas se empleará un problema de Johnson Filtration, Inc.

Un ejemplo: Johnson Filtration, Inc.

Johnson Filtration Inc. da servicio de mantenimiento a los sistemas de filtración en el sur de Florida. Los clientes llaman a Johnson Filtration, Inc. solicitando un servicio de mantenimiento para sus sistemas de filtración de agua para estimar el tiempo que se requerirá para el servicio y el

TABLA 15.5 DATOS PARA EL EJEMPLO DE JOHNSON FILTRATION

Solicitud de servicio	Meses transcurridos desde el último servicio	Tipo de reparación	Tiempo en horas necesario para la reparación
1	2	eléctrico	2.9
2	6	mecánico	3.0
3	8	eléctrico	4.8
4	3	mecánico	1.8
5	2	eléctrico	2.9
6	7	eléctrico	4.9
7	9	mecánico	4.2
8	8	mecánico	4.8
9	4	eléctrico	4.4
10	6	eléctrico	4.5

costo del mismo, los administradores de Johnson desean poder predecir este tiempo para cada solicitud de servicio. Por lo tanto, el tiempo, en horas, requerido para la reparación es la variable dependiente. Se cree que el tiempo requerido para una reparación está relacionado con dos factores, meses transcurridos desde el último servicio de mantenimiento y tipo del problema (mecánico o eléctrico). En la tabla 15.5 se presentan los datos correspondientes a una muestra de 10 solicitudes de servicio.

Sea y el tiempo, en horas, necesario para la reparación y x_1 los meses transcurridos desde el último mantenimiento. El modelo de regresión en el que sólo se usa x_1 para predecir y es

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

Empleando Minitab para obtener la ecuación de regresión estimada se tienen los resultados que se presentan en la figura 15.7. La ecuación de regresión estimada es

$$\hat{y} = 2.15 + .304x_1 \tag{15.16}$$

Empleando como nivel de significancia 0.05, el valor-p para la prueba t (o F), que es 0.016, indica que los meses transcurridos desde el último mantenimiento están relacionados significantemente con el tiempo que se requiere para la reparación. R-sq = 53.4% indica que x_1 explica sólo el 53.4% de la variabilidad en el tiempo que se requiera para una reparación.

FIGURA 15.7 RESULTADOS DE MINITAB PARA EL PROBLEMA DE JOHNSON FILTRATION EMPLEANDO COMO VARIABLE INDEPENDIENTE (x_1) EL NÚMERO DE MESES DESDE EL ÚLTIMO SERVICIO

Los nombres de las variables Months *(meses) y* Time *(tiempo) fueron ingresados en la hoja de cálculo como títulos de las columnas; por tanto,* x_1 = Months *y* y = Time.

```
The regression equation is
Time = 2.15 + 0.304 Months

Predictor     Coef   SE Coef     T      p
Constant    2.1473    0.6050  3.55  0.008
Months      0.3041    0.1004  3.03  0.016

S = 0.7810   R-sq = 53.4%   R-sq(adj) = 47.6%

Analysis of Variance

SOURCE          DF        SS       MS     F      p
Regression       1    5.5960   5.5960  9.17  0.016
Residual Error   8    4.8800   0.6100
Total            9   10.4760
```

TABLA 15.6 DATOS PARA EL EJEMPLO DE JOHNSON FILTRATION INDICANDO EL TIPO DE REPARACIÓN POR MEDIO DE UNA VARIABLE FICTICIA (x_2 = 0 PARA FALLA MECÁNICA; x_2 = 1 PARA FALLA ELÉCTRICA)



Johnson

Cliente	Meses transcurridos desde el último mantenimiento (x_1)	Tipo de reparación (x_2)	Tiempo en horas necesarias para la reparación (y)
1	2	1	2.9
2	6	0	3.0
3	8	1	4.8
4	3	0	1.8
5	2	1	2.9
6	7	1	4.9
7	9	0	4.2
8	8	0	4.8
9	4	1	4.4
10	6	1	4.5

Para incorporar en el modelo el tipo de reparación, se define la variable siguiente.

$$x_2 = \begin{cases} 0 \text{ si el tipo de reparación es mecánica} \\ 1 \text{ si el tipo de reparación es eléctrica} \end{cases}$$

En el análisis de regresión a x_2 se le llama **variable ficticia** o variable *indicadora*. Empleando esta variable ficticia, el modelo de regresión múltiple se expresa como sigue.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

En la tabla 15.6 se presentan los datos de la tabla 15.5, incluyendo los valores de la variable ficticia. Con Minitab y los datos de la tabla 15.6 se obtienen estimaciones para los parámetros del modelo. En el resultado de Minitab de la figura 15.8 se indica que la ecuación de regresión múltiple estimada es

$$\hat{y} = 0.93 + 0.388x_1 + 1.26x_2 \tag{15.17}$$

Empleando 0.05 como nivel de significancia, el valor-p correspondiente al estadístico de prueba F (F = 21.36) es 0.001, lo cual indica que la relación de regresión es significativa. En la figura 15.8, en la parte de los resultados de Minitab que corresponde a la prueba t, se observa que tanto meses transcurridos desde el último servicio (valor-p = 0.000) como tipo de reparación (valor-p = 0.005) son estadísticamente significativos. Además, R-sq = 85.9% y R-sq(adj) = 81.9% indican que la ecuación de regresión estimada explica adecuadamente la variabilidad en el tiempo necesitado para la reparación. Por lo tanto, la ecuación 15.17 sí es útil para estimar el tiempo necesario para la reparación de las diversas solicitudes de servicio.

Interpretación de los parámetros

La ecuación de regresión múltiple para el ejemplo de Johnson Filtration es

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{15.18}$$

Para entender cómo interpretar los parámetros β_0, β_1, y β_2, cuando hay una variable cualitativa, considérese el caso en que $x_2 = 0$ (reparación mecánica). Usando E (y | mecánica) para denotar la media o valor esperado del tiempo necesario para una reparación *dado* que se trata de una reparación mecánica, se tiene

$$E(y \mid \text{mecánica}) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1 \tag{15.19}$$

FIGURA 15.8 RESULTADOS DE MINITAB PARA EL EJEMPLO DE JOHNSON FILTRATION TENIENDO MESES DESDE EL ÚLTIMO SERVICIO (x_1) Y TIPO DE REPARACIÓN (x_2) COMO VARIABLES INDEPENDIENTES

Los nombres de las variables Months *(meses) y* Time *(tiempo) que aparecen en los resultados de Minitab fueron ingresados en la hoja de cálculo de Minitab como títulos de las columnas; por tanto,* x_1 = Months, x_2 = Type *(tipo) y* y = Time.

```
The regression equation is
Time = 0.930 + 0.388 Months + 1.26 Type

Predictor      Coef  SE Coef     T      p
Constant     0.9305   0.4670  1.99  0.087
Months      0.38762  0.06257  6.20  0.000
Type         1.2627   0.3141  4.02  0.005

S = 0.4590   R-sq = 85.9%   R-sq(adj) = 81.9%

Analysis of Variance

SOURCE          DF       SS      MS      F      p
Regression       2   9.0009  4.5005  21.36  0.001
Residual Error   7   1.4751  0.2107
Total            9  10.4760
```

De manera similar, en el caso de una reparación eléctrica ($x_2 = 1$), se tiene

$$E(y \mid \text{eléctrica}) = \beta_0 + \beta_1 x_1 + \beta_2(1) = \beta_0 + \beta_1 x_1 + \beta_2 \qquad \textbf{(15.20)}$$
$$= (\beta_0 + \beta_2) + \beta_1 x_1$$

Comparando las ecuaciones (15.19) y (15.20) se ve que la media del tiempo requerido para hacer una reparación es función lineal de x_1 tanto cuando se trata de reparaciones mecánicas como de reparaciones eléctricas. La pendiente en ambas ecuaciones es β_1, pero la intersección con el eje y varia. En la ecuación (15.19), para las reparaciones mecánicas, la intersección con el eje y es β_0 y en la ecuación (15.20), la ecuación para reparaciones eléctricas, la intersección es $(\beta_0 + \beta_2)$. La interpretación de β_2 es que indica la diferencia entre la media del tiempo que se requiere para una reparación eléctrica y la media del tiempo que se requiere para una reparación mecánica.

Si β_2 es positiva la media del tiempo necesario para una reparación eléctrica será mayor que para una reparación mecánica.; si β_2 es negativa la media del tiempo requerido para una reparación eléctrica será menor que para una reparación mecánica. Por último, si $\beta_2 = 0$, no hay diferencia entre las medias del tiempo que necesita para reparaciones eléctricas y mecánicas y el tipo de reparación no está relacionado con el tiempo necesario para hacer una reparación.

Empleando la ecuación de regresión múltiple estimada $\hat{y} = 0.93 + 0.388x_1 + 1.26x_2$, se ve que 0.93 es la estimación de β_0 y la estimación de β_2 es 1.26. Por lo tanto, cuando $x_2 = 0$ (reparación mecánica)

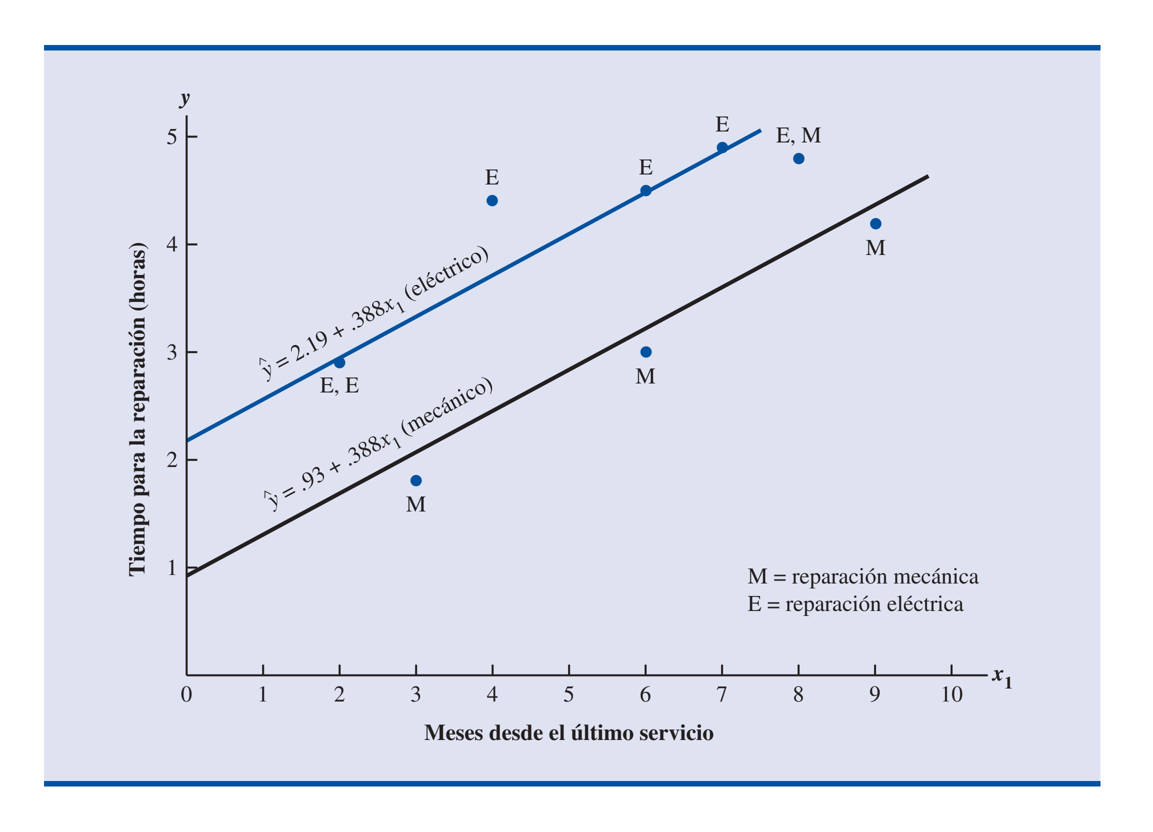
$$\hat{y} = 0.93 + 0.388x_1 \qquad \textbf{(15.21)}$$

y cuando $x_2 = 1$ (reparación eléctrica)

$$\hat{y} = 0.93 + 0.388x_1 + 1.2(1) \qquad \textbf{(15.22)}$$
$$= 2.19 + 0.388x_1$$

De esta manera, el uso de una variable ficticia proporciona dos ecuaciones que sirven para predecir el tiempo requerido para una reparación, una ecuación corresponde a las reparaciones me-

FIGURA 15.9 DIAGRAMA DE DISPERSIÓN



cánicas y la otra a las reparaciones eléctricas. Además, como $b_2 = 1.26$, se sabe que, en promedio, en las reparaciones eléctricas se necesitan 1.26 horas más que en las reparaciones mecánicas.

En la figura 15.9 se presenta una gráfica con los datos de la tabla 15.6. El tiempo de reparación, en horas, se ha representado en el eje vertical y los meses transcurridos desde el último servicio se han representado en el eje horizontal. Los puntos de la gráfica que corresponden a una reparación eléctrica se han indicado con una E y los que corresponden a una reparación mecánica con una M. En esta gráfica se han representado también las ecuaciones (15.21) y (15.22) con objeto de mostrar gráficamente las dos ecuaciones que sirven para predecir el tiempo que se requerirá para una reparación mecánica y para una reparación eléctrica.

Variables cualitativas más complejas

Para modelar una variable cualitativa que tenga k niveles se requieren k − 1 variables ficticias. Se debe ser cuidadoso al definir e interpretar variables ficticias.

En el caso del ejemplo de Johnson Filtration, como la variable cualitativa tenía dos niveles (mecánico y eléctrico), fue fácil definirla empleando cero para indicar una reparación mecánica y uno para indicar una reparación eléctrica. Sin embargo, cuando una variable cualitativa tiene más de dos niveles, habrá que tener cuidado tanto al definir como al interpretar estas variables ficticias. Como se verá a continuación, si una variable ficticia tiene k niveles, se necesitan $k - 1$ variables ficticias, cada una de las cuales tomará el valor 0 o 1.

Supóngase, por ejemplo, que un fabricante de fotocopiadoras divide un estado en tres regiones de ventas A, B y C. Sus gerentes desean emplear el análisis de regresión para predecir las ventas semanales. Empleando como variable dependiente el número de fotocopiadoras vendidas, están considerando varias variables independientes (el número de vendedores, gastos en publicidad, etc.). Supóngase que los gerentes piensan que la región de ventas puede ser también un factor importante en la predicción del número de fotocopiadoras vendidas. Como región es una

variable cualitativa que tiene tres niveles, A, B y C, para representar la región de ventas se necesitarán $3 - 1 = 2$ variables ficticias. Cada variable tomará los valores 0 o 1.

$$x_1 = \begin{cases} 1 \text{ si la región de ventas es B} \\ 0 \text{ si no es así} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ si la región de ventas es C} \\ 0 \text{ si no es así} \end{cases}$$

De acuerdo con esta definición, para x_1 y x_2 se tienen los valores siguientes.

Región	x_1	x_2
A	0	0
B	1	0
C	0	1

En las observaciones correspondientes a la región A se tendrá $x_1 = 0$, $x_2 = 0$; en las observaciones correspondientes a la región B se tendrá $x_1 = 1$, $x_2 = 0$, y en las observaciones correspondientes a la región C se tendrá $x_1 = 0$, $x_2 = 1$.

La ecuación de regresión que relaciona el valor esperado del número de fotocopiadoras vendidas $E(y)$, con las variables ficticias se expresa como sigue.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Para interpretar los parámetros β_0, β_1, y β_2, considérense las siguientes tres variaciones de la ecuación de regresión.

$$E(y \mid \text{región A}) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$
$$E(y \mid \text{región B}) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$
$$E(y \mid \text{región C}) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

Por lo tanto, β_0, es la media o valor esperado de las ventas en la región A; β_1 es la diferencia entre la media del número de unidades vendidas en la región B y la media del número de unidades vendidas en la región A, y β_2 es la diferencia entre la media del número de unidades vendidas en la región C y la media del número de unidades vendidas en la región A.

Se necesitaron dos variables ficticias debido a que la región de ventas es una variable cualitativa que tiene tres niveles. Sin embargo, la asignación que se hizo al usar $x_1 = 0$, $x_2 = 0$ para identificar la región A, $x_1 = 1$, $x_2 = 0$ para identificar la región B y $x_1 = 0$, $x_2 = 1$ para identificar la región C fue arbitraria. De igual manera se podría haber elegido, por ejemplo, $x_1 = 1$, $x_2 = 0$ para identificar la región A, $x_1 = 0$, $x_2 = 0$ para identificar la región B y $x_1 = 0$, $x_2 = 1$ para identificar la región C. En ese caso, β_1 se habría interpretado como la media de la diferencia entre las regiones A y B, y β_2 como la media de la diferencia entre las regiones C y B.

Es importante recordar que en el análisis de regresión múltiple, cuando una variable cualitativa tiene k niveles, se requieren $k - 1$ variables ficticias. Entonces, si en el ejemplo de las regiones de ventas hubiera una cuarta región, D, se necesitarían tres variables ficticias. Por ejemplo las tres variables ficticias

$$x_1 = \begin{cases} 1 \text{ si la región es B} \\ 0 \text{ si no es así} \end{cases} \quad x_2 = \begin{cases} 1 \text{ si la región es C} \\ 0 \text{ si no es así} \end{cases} \quad x_3 = \begin{cases} 1 \text{ si la región es D} \\ 0 \text{ si no es así} \end{cases}$$

Ejercicios

Métodos



32. Considere un estudio de regresión en el que intervienen una variable dependiente y, una variable independiente cuantitativa x_1 y una variable cualitativa de dos niveles (nivel 1 y nivel 2).
    a. Dé la ecuación de regresión múltiple que relaciona x_1 y la variable cualitativa con y.
    b. ¿Cuál es el valor esperado de y que corresponde al nivel 1 de la variable cualitativa?
    c. ¿Cuál es el valor esperado de y que corresponde al nivel 2 de la variable cualitativa?
    d. Interprete los parámetros de la ecuación de regresión.
33. Considere un estudio de regresión en el que intervienen una variable dependiente y, una variable independiente cuantitativa x_1 y una variable cualitativa de tres niveles (nivel 1, nivel 2 y nivel 3).
    a. ¿Cuántas variables ficticias se requieren para representar la variable cualitativa?
    b. Dé una ecuación de regresión múltiple que relacione x_1 y la variable cualitativa con y.
    c. Interprete los parámetros de la ecuación de regresión.

Aplicaciones



34. El administrador propuso el siguiente modelo de regresión para predecir las ventas en un punto de venta de comida rápida.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

donde

$$x_1 = \text{número de competidores a no más de una milla}$$
$$x_2 = \text{población a no más de una milla (en miles)}$$
$$x_3 = \begin{cases} 1 \text{ si tiene ventanilla para conductores} \\ 0 \text{ si no es así} \end{cases}$$
$$y = \text{ventas (en miles de \$)}$$

Se obtuvo la siguiente ecuación de regresión estimada con los datos de 20 puntos de venta.

$$\hat{y} = 10.1 - 4.2x_1 + 6.8x_2 + 15.3x_3$$

a. ¿Cuál es la cantidad esperada de ventas atribuible a la ventana para conductores?
b. Pronostique las ventas de un negocio que tiene dos competidores y una población de 8000 a no más de una milla y ventana para los conductores.
c. Pronostique las ventas de un negocio que tiene un competidor y una población de 3000 a no más de una milla y ventana para los conductores.

35. Véase el problema de Johnson Filtration presentado en esta sección. Supóngase que además de la información sobre los meses transcurridos desde el último servicio y de si se trata de una reparación mecánica o eléctrica, los administradores presentan una lista con las personas que realizaron las reparaciones. A continuación se presentan los nuevos datos.



Tiempo en horas requerido para la reparación	Meses desde el último servicio	Tipo de reparación	Persona que realiza la reparación
2.9	2	eléctrica	Dave Newton
3.0	6	mecánica	Dave Newton
4.8	8	eléctrica	Bob Jones

(*continúa*)

Tiempo en horas requerido para la reparación	Meses desde el último servicio	Tipo de reparación	Persona que realiza la reparación
1.8	3	mecánica	Dave Newton
2.9	2	eléctrica	Dave Newton
4.9	7	eléctrica	Bob Jones
4.2	9	mecánica	Bob Jones
4.8	8	mecánica	Bob Jones
4.4	4	eléctrica	Bob Jones
4.5	6	eléctrica	Dave Newton

a. Por ahora ignore los meses transcurridos desde el último servicio (x_1) y la persona que realizó la reparación. Obtenga la ecuación de regresión lineal simple estimada para predecir el tiempo que se requiere para la reparación (y) dado el tipo de reparación (x_2). Recuerde que $x_2 = 0$ si se trata de una reparación mecánica, y $x_2 = 1$ si se trata de una reparación eléctrica.
b. ¿Proporciona la ecuación obtenida en el inciso a) un buen ajuste a los datos observados? Explique.
c. Por ahora ignore los meses transcurridos desde el último servicio y el tipo de reparación. Obtenga la ecuación de regresión lineal simple estimada para predecir el tiempo necesario para la reparación dada la persona que realizó la reparación. Sea $x_3 = 0$ si la reparación fue hecha por Bob Jones, y $x_3 = 1$ si la reparación fue hecha por Dave Newton.
d. ¿Proporciona la ecuación obtenida en el inciso c) un buen ajuste a los datos observados? Explique.

36. Este problema es una extensión del ejercicio 35.
    a. Obtenga la ecuación de regresión estimada que permita predecir el tiempo que se requiere para una reparación dados los meses transcurridos desde la última reparación, el tipo de reparación y la persona que realizó la reparación.
    b. Empleando como nivel de significancia 0.05, realice una prueba para ver si la ecuación de regresión estimada obtenida en el inciso a) representa una relación significativa entre las variables independientes y la variable dependiente.
    c. ¿Es estadísticamente significativo agregar la variable x_3, la persona que realizó la reparación? Use $\alpha = 0.05$. ¿Qué explicación puede dar para los resultados observados?

37. La Liga nacional de futbol americano de Estados Unidos (National Football League) evalúa a sus prospectos con una escala que va del 5 al 9. Estas evaluaciones se interpretan como sigue: 8 – 9 deberá empezar el año próximo; 7.0 – 7.9 deberá empezar; 6.0 – 6.9 servirán de respaldo al equipo, y 5.0 – 5.9 pueden formar parte del club y contribuir. En la tabla siguiente se da posición, peso, tiempo en segundos para correr 40 yardas y la evaluación dada por la NFL a 25 prospectos (*USA Today*, 14 de abril de 2000).
    a. Dé una variable ficticia para la posición de los jugadores.



Football

	Posición	Peso (libras)	Tiempo (segundos)	Evaluación
Cosey Coleman	Guardia	322	5.38	7.4
Travis Claridge	Guardia	303	5.18	7.0
Kaulana Noa	Guardia	317	5.34	6.8
Leander Jordan	Guardia	330	5.46	6.7
Chad Clifton	Guardia	334	5.18	6.3
Manula Savea	Guardia	308	5.32	6.1
Ryan Johanningmeir	Guardia	310	5.28	6.0
Mark Tauscher	Guardia	318	5.37	6.0
Blaine Saipaia	Guardia	321	5.25	6.0
Richard Mercier	Guardia	295	5.34	5.8
Damion McIntosh	Guardia	328	5.31	5.3

	Posición	Peso (libras)	Tiempo (segundos)	Evaluación
Jeno James	Guardia	320	5.64	5.0
Al Jackson	Guardia	304	5.20	5.0
Chris Samuels	Tackle	325	4.95	8.5
Stockar McDougle	Tackle	361	5.50	8.0
Chris McIngosh	Tackle	315	5.39	7.8
Adrian Klemm	Tackle	307	4.98	7.6
Todd Wade	Tackle	326	5.20	7.3
Marvel Smith	Tackle	320	5.36	7.1
Michael Thompson	Tackle	287	5.05	6.8
Bobby Williams	Tackle	332	5.26	6.8
Darnell Alford	Tackle	334	5.55	6.4
Terrance Beadles	Tackle	312	5.15	6.3
Tutan Reyes	Tackle	299	5.35	6.1
Greg Robinson-Ran	Tackle	333	5.59	6.0

b. Obtenga una ecuación de regresión estimada que muestre la relación entre la evaluación y posición, peso y tiempo requerido para correr 40 yardas.

c. Empleando como nivel de significancia 0.05, pruebe si la ecuación de regresión estimada obtenida en el inciso b) indica que existe una relación significativa entre las variables independientes y la variable dependiente.

d. ¿Proporciona la ecuación de regresión estimada un buen ajuste a los datos observados? Explique.

e. ¿Es la posición un factor significativo en la evaluación de los jugadores? Use $\alpha = 0.05$. Explique.

f. Suponga que hay un nuevo prospecto de tackle que pesa 300 libras y corre 40 yardas en 5.1 segundos. Utilice la ecuación de regresión estimada obtenida en el inciso b) para estimar la evaluación de este jugador.

38. Un estudio realizado a lo largo de 10 años por la American Heart Association proporcionó datos sobre la relación que tienen la edad, la presión sanguínea y el fumar sobre el riesgo de sufrir un infarto. Los datos que se dan a continuación se obtuvieron como parte de este estudio. El riesgo se interpreta como la probabilidad (multiplicada por 100) de que el paciente sufra un infarto en los próximos 10 años. Para fumar, defina una variable ficticia que tome el valor 1 si la persona es fumadora y el valor 0 si no es fumadora.



Stroke

Riesgo	Edad	Presión	Fumador
12	57	152	No
24	67	163	No
13	58	155	No
56	86	177	Sí
28	59	196	No
51	76	189	Sí
18	56	155	Sí
31	78	120	No
37	80	135	Sí
15	78	98	No
22	71	152	No
36	70	173	Sí
15	67	135	Sí
48	77	209	Sí
15	60	199	No
36	82	119	Sí
8	66	166	No
34	80	125	Sí
3	62	117	No
37	59	207	Sí

a. Obtenga la ecuación de regresión estimada que relaciona el riesgo de infarto con la edad, la presión sanguínea y el fumar o no fumar.
b. ¿Es el fumar un factor significativo para el riesgo de infarto? Explique. Use $\alpha = 0.05$.
c. ¿Cuál es la probabilidad de que Art Apeen sufra un infarto en los próximos 10 años, si tiene 68 años, fuma y su presión sanguínea es 175? ¿Qué recomendará el médico hacer a este paciente?



15.8 Análisis residual

En el capítulo 14 se indicó que los residuales estandarizados suelen emplearse en las gráficas de residuales y en la identificación de observaciones atípicas. A continuación se presenta la fórmula general para obtener el residual estandarizado de la observación i.

RESIDUAL ESTANDARIZADO DE LA OBSERVACIÓN i

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{15.23}$$

donde

$$s_{y_i - \hat{y}_i} = \text{desviación estándar del residual } i$$

La fórmula general para obtener la desviación estándar del residual i está definida como se indica a continuación.

DESVIACIÓN ESTÁNDAR DEL RESIDUAL i

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \tag{15.24}$$

donde

$$s = \text{error estándar de estimación}$$
$$h_i = \text{influencia de la observación } i$$

Como se dijo en el capítulo 14, la **influencia** de una observación está determinada por qué tan lejos de sus medias están los valores de las variables independientes. En el análisis de regresión múltiple, calcular h_i y $s_{y_i - \hat{y}_i}$, y por lo tanto el residual estandarizado de la observación i es muy complicado como para hacerlo a mano. Sin embargo, los residuales estandarizados se obtienen fácilmente como parte de los resultados de los paquetes de software para estadística. En la tabla 15.7 se presentan valores pronosticados, residuales y residuales estandarizados empleando los datos del ejemplo de Butler Trucking presentado previamente en este capítulo; estos valores se obtuvieron empleando Minitab. Los valores pronosticados que aparecen en la tabla están basados en la ecuación de regresión estimada $\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$.

Los residuales estandarizados y los valores pronosticados de y de la tabla 15.7 se emplearon en la figura 15.10, la gráfica de residuales estandarizados para el ejemplo de regresión múltiple de Butler Trucking. En esta gráfica de residuales estandarizados no se observa ninguna anormalidad. Además, todos los residuales estandarizados se encuentran entre -2 y $+2$; por lo tanto no hay ninguna razón para cuestionar la suposición de que el término del error esté distribuido normalmente. Así, se concluye que las suposiciones del modelo son razonables.

TABLA 15.7 RESIDUALES Y RESIDUALES ESTANDARIZADOS CORRESPONDIENTES AL ANÁLISIS DE REGRESIÓN DE BUTLER TRUCKING.

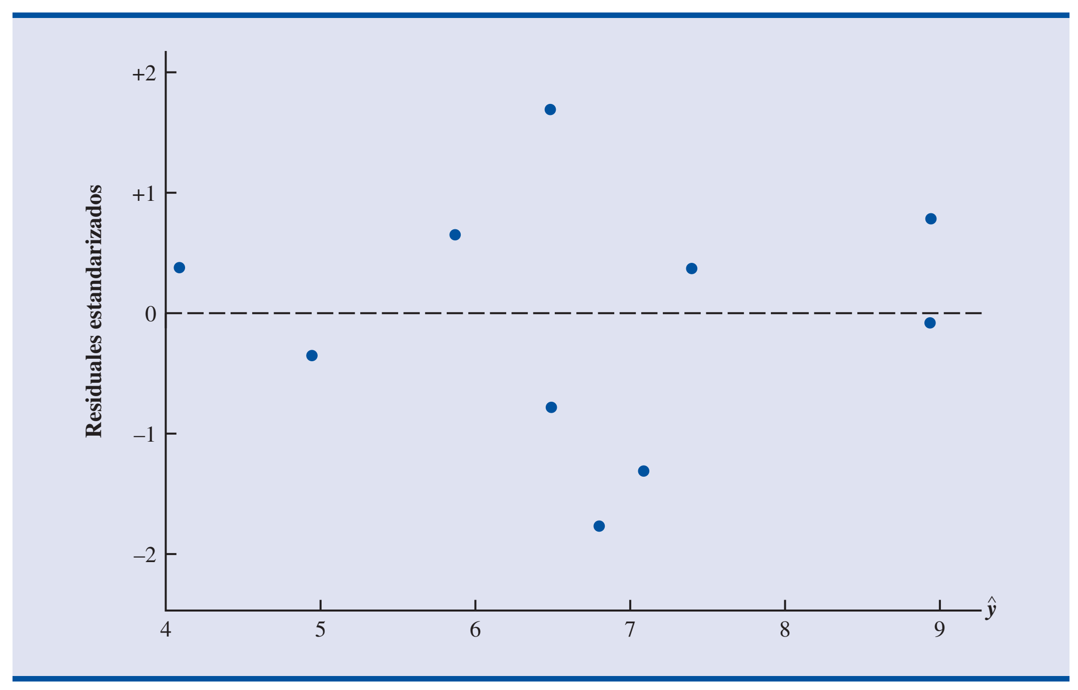
Millas recorridas (x_1)	Entregas (x_2)	Tiempo de recorrido (y)	Valor pronosticado ($\hat{y}$)	Residual ($y - \hat{y}$)	Residual estandarizado
100	4	9.3	8.93846	0.361541	0.78344
50	3	4.8	4.95830	−0.158304	−0.34962
100	4	8.9	8.93846	−0.038460	−0.08334
100	2	6.5	7.09161	−0.591609	−1.30929
50	2	4.2	4.03488	0.165121	0.38167
80	2	6.2	5.86892	0.331083	0.65431
75	3	7.4	6.48667	0.913331	1.68917
65	4	6.0	6.79875	−0.798749	−1.77372
90	3	7.6	7.40369	0.196311	0.36703
90	2	6.1	6.48026	−0.380263	−0.77639

Para determinar si la distribución de ϵ parece ser normal, puede emplearse también una gráfica de probabilidad normal. En la sección 14.8 se discutió el procedimiento y la interpretación de una gráfica de probabilidad normal. Ese mismo procedimiento es adecuado para la regresión múltiple. Para obtener la grafica de probabilidad normal también se hace uso de un paquete de software para estadística que realice los cálculos.

Detección de observaciones atípicas

Una **observación atípica** es una observación que es inusual en relación con el resto de los datos; en otras palabras, una observación atípica no sigue el patrón del resto de los datos. En el capítulo 14 se mostró un ejemplo en el que había una observación atípica y se vio el empleo de los residuales estandarizados para detectar observaciones atípicas. Minitab clasifica una observación

FIGURA 15.10 GRÁFICA DE RESIDUALES ESTANDARIZADOS EMPLEANDO EL EJEMPLO DE BUTLER TRUCKING



como observación atípica si el valor de su residual estandarizado es menor a −2 o mayor a +2. Aplicando esta regla a los residuales estandarizados del ejemplo de Butler Trucking (tabla 15.7), en este conjunto de datos no se detecta ninguna observación atípica.

En general, la presencia de una o más observaciones atípicas en un conjunto de datos tiende a incrementar s, el error estándar de estimación y, por lo tanto, a incrementar $s_{y - \hat{y}_i}$, la desviación estándar del residual i. Dado que $s_{y_i - \hat{y}_i}$ aparece como denominador en la fórmula (15.23) del residual estandarizado, el tamaño del residual estandarizado disminuirá a medida que s aumente. Esto da como resultado que aún cuando un residual sea inusualmente grande, el denominador de la fórmula (15.23), que será grande, hará que la regla del residual estandarizado falle para la identificación de una observación como observación atípica. Es posible sortear esta dificultad empleando una forma de los residuales estandarizados conocida como **residuales estudentizados**.

Residuales estudentizados eliminados y observaciones atípicas

Supóngase que del conjunto de datos se elimina la observación i y que de las $n - 1$ observaciones restantes se obtiene una nueva ecuación de regresión estimada. Sea $s_{(i)}$ el error estándar de estimación obtenido del conjunto de datos en los que se ha eliminado la observación i. Si se calcula la desviación estándar del residual i usando $s_{(i)}$ en lugar de s y después se calcula el residual estandarizado de la observación i empleando el nuevo valor de $s_{y_i - \hat{y}_i}$, al residual estandarizado que se obtiene se le llama residual eliminado estudentizado. Si la observación i es una observación atípica, $s_{(i)}$ será menor a s. Por lo tanto, el valor absoluto del residual eliminado estudentizado i será mayor que el valor absoluto del residual estandarizado. De esta manera, los residuales eliminados estudentizados pueden detectar observaciones atípicas que los residuales estandarizados no detectan.

Muchos de los paquetes de software para estadística proporcionan una opción para obtener residuales eliminados estudentizados. Empleando Minitab se obtuvieron los residuales eliminados estudentizados para el ejemplo de Butler Trucking; los resultados obtenidos se presentan en la tabla 15.8. Para determinar si los residuales eliminados estudentizados indican la presencia de observaciones atípicas se emplea la distribución t. Recuérdese que p denota el número de variables independientes y n el número de observaciones. Por lo tanto, si se elimina la observación i, el número de observaciones en el nuevo conjunto de datos es $n - 1$; en este caso, la suma de cuadrados del error tiene $(n - 1) - p - 1$ grados de libertad. Como en el ejemplo de Butler Trucking $n = 10$ y $p = 2$, los grados de libertad para la suma de cuadrados del error es $9 - 2 - 1 = 6$. Empleando como nivel de significancia 0.05, en la distribución t (tabla 2 del apéndice B) para seis grados de libertad se obtiene, $t_{0.025} = 2.337$. Se concluye que la observación i es una observación atípica si el residual eliminado estudentizado es menor a -2.447 o mayor a $+2.447$. En la tabla 15.8 se observa que los residuales eliminados estudentizados no se encuentran fuera de estos límites; por lo tanto se concluye que en este conjunto de datos no hay observaciones atípicas.

TABLA 15.8 RESIDUALES ELIMINADOS ESTUDENTIZADOS CORRESPONDIENTES AL EJEMPLO DE BUTLER TRUCKING

Millas recorridas (x_1)	Entregas (x_2)	Tiempo recorrido (y)	Residual estandarizado	Residual eliminado estudentizado
100	4	9.3	0.78344	0.75939
50	3	4.8	−0.34962	−0.32654
100	4	8.9	−0.08334	−0.07720
100	2	6.5	−1.30929	−1.39494
50	2	4.2	0.38167	0.35709
80	2	6.2	0.65431	0.62519
75	3	7.4	1.68917	2.03187
65	4	6.0	−1.77372	−2.21314
90	3	7.6	0.36703	0.34312
90	2	6.1	−0.77639	−0.75190

TABLA 15.9 INFLUENCIA Y DISTANCIA DE COOK CORRESPONDIENTES AL EJEMPLO DE BUTLER TRUCKING

Millas recorridas (x_1)	Entregas (x_2)	Tiempo de recorrido (y)	Influencia (h_i)	D Cook (D_i)
100	4	9.3	0.351704	0.110994
50	3	4.8	0.375863	0.024536
100	4	8.9	0.351704	0.001256
100	2	6.5	0.378451	0.347923
50	2	4.2	0.430220	0.036663
80	2	6.2	0.220557	0.040381
75	3	7.4	0.110009	0.117562
65	4	6.0	0.382657	0.650029
90	3	7.6	0.129098	0.006656
90	2	6.1	0.269737	0.074217

Observaciones influyentes

En la sección 14.9 se vio cómo se puede usar la influencia de una observación para identificar observaciones cuyo valor de la variable independiente puede tener una influencia fuerte en los resultados de la regresión. Como se indicó al hablar de los residuales estandarizados, la influencia de una observación, que se denota h_i, mide qué tan lejos de sus medias se encuentran los valores de las variables independientes. Los valores de la influencia se pueden obtener como parte de los resultados que proporcionan los paquetes de software para estadística. Minitab calcula los valores de la influencia y para detectar **observaciones influyentes** emplea la regla $h_i > 3(p + 1)/n$. En el ejemplo de Butler Trucking como hay $p = 2$ variables independientes y $n = 10$ observaciones, el valor crítico para la influencia es $3(2 + 1)/10 = 0.9$. En la tabla 15.9 se presentan los valores de la influencia correspondientes al ejemplo de Butler Trucking obtenidos con Minitab. Como ninguno de los valores h_i es mayor a 0.9, en este conjunto de datos no se detectan observaciones influyentes.

Uso de la medida de la distancia de Cook para identificar observaciones influyentes

Un problema que puede presentarse al usar la influencia para identificar observaciones influyentes es que puede que se identifique una observación como una observación que tiene una gran influencia sin que necesariamente sea influyente en términos de la ecuación de regresión estimada que se obtiene. Por ejemplo, en la tabla 15.10 se presenta un conjunto de datos que consta de 10 observaciones y sus correspondientes valores de influencia (obtenidos usando Minitab). Como la influencia de la última observación es $0.91 > 0.75$ (el valor de influencia crítico), se identificará a esta observación como una observación influyente. Sin embargo antes de aceptar una conclusión final, considérese la situación desde una perspectiva diferente.

TABLA 15.10 CONJUNTO DE DATOS QUE ILUSTRAN EL PROBLEMA POTENCIAL QUE EXISTE USANDO EL CRITERIO DE LA INFLUENCIA

x_i	y_i	Influencia h_i
1	18	0.204170
1	21	0.204170
2	22	0.164205
3	21	0.138141
4	23	0.125977
4	24	0.125977
5	26	0.127715
15	39	0.909644

En la figura 15.11 se presenta el diagrama de dispersión que corresponde al conjunto de datos de la tabla 15.10. Empleando Minitab se obtuvo a partir de estos datos la ecuación de regresión estimada siguiente.

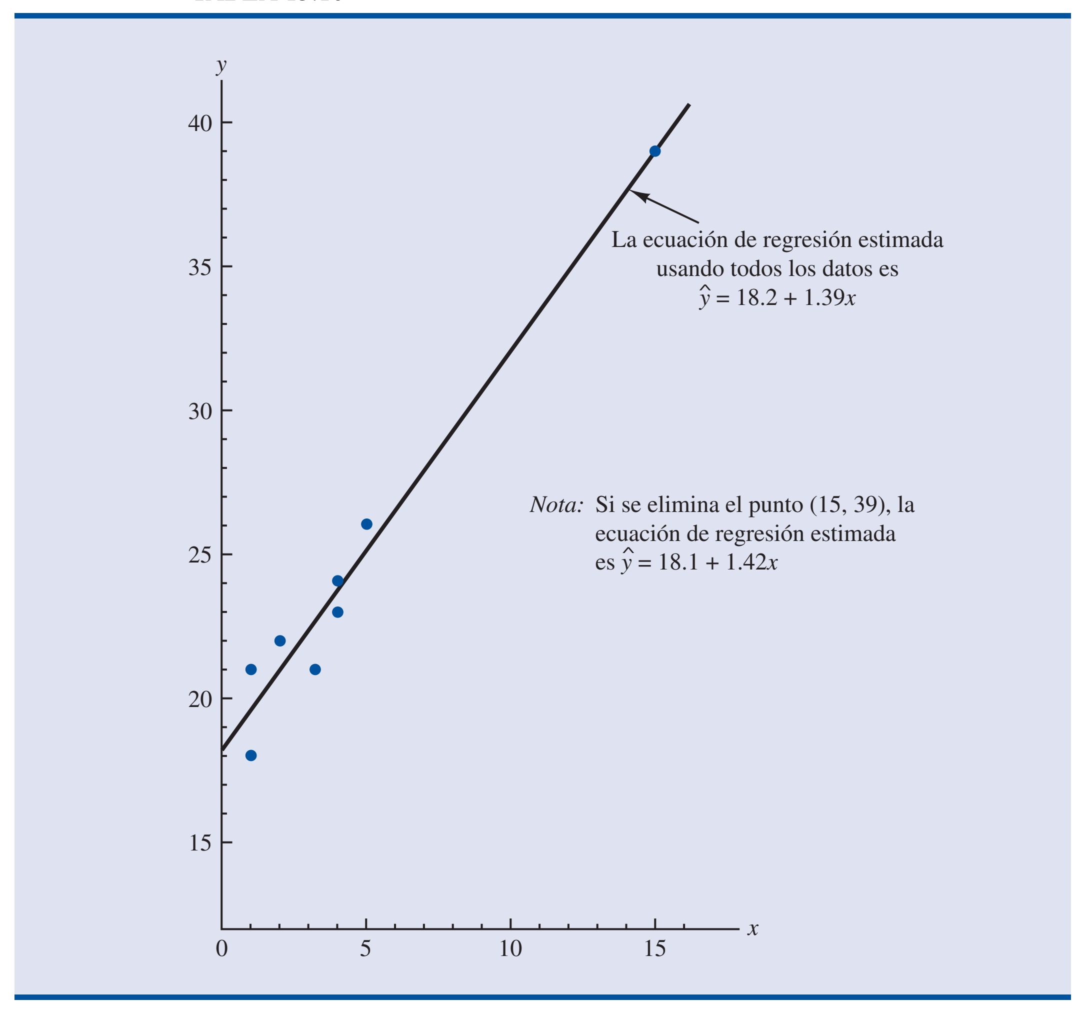
$$\hat{y} = 18.2 + 1.39x$$

La línea recta que se observa en la figura 15.11 es la gráfica de esta ecuación. Ahora, si de este conjunto de datos se elimina la observación $x = 15$, $y = 39$ y con las siete observaciones restantes se obtiene una nueva ecuación, la nueva ecuación de regresión estimada es

$$\hat{y} = 18.1 + 1.42x$$

Se observa que en la nueva ecuación la intersección con el eje y y la pendiente no tienen valores significativamente diferentes a los de la ecuación en la que se usan todos los datos. A pesar de

FIGURA 15.11 DIAGRAMA DE DISPERSIÓN OBTENIDO CON LOS DATOS DE LA TABLA 15.10



que con el criterio de la influencia se identificó a la octava observación con observación influyente, es claro que esta observación tiene poca influencia en los resultados. Por lo tanto, hay casos en los que emplear únicamente la influencia para identificar las observaciones influyentes puede llevar a conclusiones erróneas.

En la **medida de la distancia de Cook** se utiliza tanto la influencia de la observación i, h_i, como el residual de la observación i, $(y_i - \hat{y}_i)$, para determinar si una observación es influyente.

MEDIDA DE LA DISTANCIA DE COOK

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(p - 1)s^2}\left[\frac{h_i}{(1 - h_i)^2}\right] \qquad \textbf{(15.25)}$$

donde

D_i = medida de la distancia de Cook para la observación i
$y_i - \hat{y}_i$ = residual de la observación i
h_i = influencia de la observación i
p = número de variables independientes
s = error estándar de estimación

Si el residual o la influencia es grande, la medida de la distancia de Cook será grande e identificará una observación influyente. Como regla general se acepta que si $D_i > 1$ la observación i es influyente y debe ser analizada con más detenimiento. En la última columna de la tabla 15.9 se presentan las medidas de la distancia de Cook correspondientes al problema Butler Trucking obtenidas con Minitab. La observación con mayor influencia es la observación 8, para la que $D_i = 0.650029$. Aplicando la regla $D_i > 1$ se ve que no hay por qué preocuparse acerca de la presencia de observaciones influyentes en el conjunto de datos de Butler Trucking.

NOTAS Y COMENTARIOS

1. Los procedimientos para la detección de observaciones atípicas o de observaciones influyentes permiten estar alerta acerca de los efectos potenciales que algunas observaciones puedan tener en los resultados de la regresión. Cada observación atípica u observación influyente justifica un examen cuidadoso. Si se encuentran errores en los datos, se pueden corregir los errores y repetir el análisis de regresión. En general, las observaciones atípicas y las observaciones influyentes no deben ser eliminadas del conjunto de datos a menos que haya una evidencia clara que indique que no provienen de elementos de la población en estudio y que no tenían que ser incluidos en el conjunto de datos original

2. Para determinar si el valor de una medida de la distancia de Cook D_i es lo suficientemente grande como para concluir que la observación i es influyente, también puede compararse el valor de D_i con el percentil 50 de una distribución F (denotado $F_{0.50}$) con $p + 1$ grados de libertad en el numerador y $n - p - 1$ grados de libertad en el denominador. Para esta prueba se necesita contar con tablas F a un nivel de significancia de 0.50. La regla práctica dada antes ($D_i > 1$) se basa en el hecho de que en muchos de los casos los valores en la tabla son cercanos a 1.

Ejercicios

Métodos



39. A continuación se dan datos para las variables x y y.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Obtenga una ecuación de regresión estimada para estos datos.
b. Grafique los residuales estandarizados contra $\hat{y}$. ¿Parece haber alguna observación atípica en este conjunto de datos? Explique.
c. Calcule los residuales eliminados estudentizados de estos datos. Empleando como nivel de significancia 0.05, ¿puede clasificarse cualquiera de estas observaciones como observación atípica? Explique.

40. A continuación se dan datos para las variables x y y.

x_i	22	24	26	28	40
y_i	12	21	31	35	70

a. Obtenga una ecuación de regresión estimada para estos datos.
b. Calcule los residuales eliminados estudentizados de estos datos Empleando como nivel de significancia 0.05, ¿puede clasificarse cualquiera de estas observaciones como observación atípica? Explique.
c. Calcule los valores de influencia de estos datos. ¿Parece haber alguna observación influyente en estos datos? Explique.
d. Calcule la medida de la distancia de Cook de estos datos. ¿Es alguna de las observaciones una observación influyente? Explique.

Aplicaciones



41. En el ejercicio 5 se presentaron los datos siguientes sobre el ingreso semanal bruto y publicidad tanto en televisión como en periódicos de Showtime Movie Theater.


Showtime

Ingreso semanal bruto (en miles de $)	Publicidad en televisión (en miles $)	Publicidad en periódicos (en miles de $)
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

a. Dé una ecuación de regresión estimada que relacione el ingreso semanal bruto con los gastos en publicidad en televisión y periódicos.
b. Grafique los residuales estandarizados contra $\hat{y}$. ¿Respalda la gráfica de residuales las suposiciones acerca de ϵ? Explique.
c. Revise que no haya observaciones atípicas en estos datos. ¿A qué conclusión llega?
d. ¿Hay alguna observación influyente?

42. En los datos siguientes se presenta peso en vacío, caballos de fuerza y velocidad en ¼ de milla de 10 automóviles deportivos y GT. Supóngase que se tiene también el precio de cada uno de estos automóviles. Todo el conjunto de datos es el siguiente.


Auto2

Automóvil deportivo y GT	Precio (miles de $)	Peso en vacío (lb)	Caballos de fueza (lb)	Velocidad en ¼ de milla (mph)
Accura Integra Type R	25 035	2577	195	90.7
Accura NSX-T	93 758	3066	290	108.0
BMW Z3 2.8	40 900	2844	189	93.2
Chevrolet Camaro Z28	24 865	3439	305	103.2
Chevrolet Corvette Convertible	50 144	3246	345	102.1
Dodge Viper RT/10	69 742	3319	450	116.2
Ford Mustang GT	23 200	3227	225	91.7
Honda Prelude Type SH	26 382	3042	195	89.7
Mercedes-Benz CLK320	44 988	3240	215	93.0
Mercedes-Benz SLK230	42 762	3025	185	92.3
Mitsubishi 3000GT VR-4	47 518	3737	320	99.0
Nissan 240SX SE	25 066	2862	155	84.6
Pontiac Firebird Trans Am	27 770	3455	305	103.2
Porsche Boxster	45 560	2822	201	93.2
Toyota Supra Turbo	40 989	3505	320	105.0
Volvo C70	41 120	3285	236	97.0

a. Obtenga la ecuación de regresión estimada en la que se emplee precio y caballos de fuerza para predecir la velocidad en ¼ de milla.
b. Grafique los residuales estandarizados contra $\hat{y}$. ¿Respalda la gráfica de residuales las suposiciones respecto a ϵ? Explique.
c. Verifique si hay observaciones atípicas. ¿A qué conclusión llega?
d. ¿Hay alguna observación influyente? Explique.



LPGA

43. La Ladies Professional Golfers Association (LPGA) lleva estadísticas sobre el desempeño y los ingresos de sus miembros en el LPGA Tour. En el archivo titulado LPGA del disco compacto se presentan las estadísticas de fin de año sobre el desempeño de las 30 jugadoras que tuvieron los mejores ingresos en el LPGA Tour del 2005 (www.lpga.com, 2006). Earnings (ingresos) ($1000) son los ingresos totales en miles de dólares; Scoring Avg es el n promedio de golpes en todo el evento; Greens in Reg. es el porcentaje de las veces que un jugador logra un *green* en regulación, y Putting Avg es el promedio de *putts* por *green* en regulación. Un *green* se considera en regulación cuando se alcanza en dos golpes menos que el par del hoyo
    a. Obtenga una ecuación de regresión estimada que sirva para obtener Scoring Avg conociendo Greens en Reg. y Putting Avr.
    b. Grafique los residuales estandarizados contra $\hat{y}$. ¿Confirma esta gráfica de residuales las suposiciones hechas acerca de ϵ?
    c. Verifique si existen observaciones atípicas. ¿A qué conclusión llega?
    d. ¿Hay alguna observación influyente? Explique.



Regresión logística

En muchas de las aplicaciones de la regresión la variable dependiente asume sólo dos valores discretos. Por ejemplo, en un banco suele necesitarse una ecuación de regresión estimada para predecir si a una persona se le aprobará su solicitud de tarjeta de crédito. A esta variable dependiente pueden dársele los valores $y = 1$ si la solicitud de tarjeta de crédito es aprobada, y $y = 0$ si es rechazada. Con la regresión logística, dado un conjunto particular de valores de las variables independientes elegidas, se estima la probabilidad de que el banco apruebe la solicitud de tarjeta de crédito.

A continuación se considera una aplicación de la regresión logística. La empresa Simmons Stores va a realizar una promoción por correo. Simmons Stores es una cadena nacional de ropa para dama. Ha mandado imprimir 5000 costosos catálogos de venta a cuatro colores y en cada catálogo incluye un cupón de $50 de descuento en la compra de $200 o más.

Como este catálogo es costoso, Simmons desea enviarlo sólo a aquellos clientes que tengan mayor probabilidad de usar el cupón. Los gerentes consideran que la cantidad gastada anualmente por el cliente en las tiendas Simmons, así como si posee o no una tarjeta de crédito de Simmons son dos variables que pueden servir para predecir si ese cliente usará el cupón. Simmons realiza un estudio piloto usando una muestra aleatoria de 50 clientes con tarjeta de crédito de Simmons y 50 clientes sin tarjeta de crédito de la empresa. Simmons envió los catálogos a cada uno de estos 100 clientes elegidos. Al final del periodo de prueba, Simmons anota si los clientes han hecho uso o no del cupón. En la tabla 15.11 se presentan los datos muestrales de las 10 primeras personas que recibieron el catálogo. En esta tabla se da: cantidad, en miles de dólares, gastada por el cliente en las tiendas Simmons durante el año anterior; información sobre si tiene o no tarjeta de crédito de Simmons, codificada como 1 si el cliente tiene tarjeta de crédito de Simmons y 0 si no la tiene. En la columna correspondiente al cupón, 1 significa el cliente usó el cupón y 0 significa no lo usó.

Para ayudar a Simmons a predecir si las personas que reciban el catálogo usarán o no el cupón, se podría pensar en construir, con los datos de la tabla 15.11, un modelo de regresión múltiple. Las variables independientes serían cantidad gastada anualmente en Simmons Stores y tarjeta de crédito, y la variable dependiente sería cupón. Sin embargo, el modelo común de regresión múltiple no se puede emplear debido a que la variable dependiente sólo puede tomar los valores 0 y 1. Con este ejemplo se ilustra el tipo de situación para la cual fue creada la regresión logística. A continuación se verá el empleo de la regresión logística para ayudar a Simmons Stores a pronosticar qué tipo de clientes es más probable que aprovechen esta promoción.

TABLA 15.11 DATOS MUESTRALES DE SIMMONS STORES


Simmons

Cliente	Gastos anual (miles de $)	Tarjeta de Simmons	Cupón
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

Ecuación de regresión logística

La regresión logística se parece en muchos aspectos a la regresión común. Se necesita una variable dependiente *y* y una o varias variables independientes. En el análisis de regresión múltiple, a la media o valor esperado de *y* se le conoce como la ecuación de regresión múltiple.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad \textbf{(15.26)}$$

En la regresión logística, tanto la teoría como la práctica estadística han demostrado que la relación existente entre $E(y)$ y $x_1, x_2, \ldots, x_p$, queda descrita mediante la siguiente ecuación no lineal.

ECUACIÓN DE REGRESIÓN LOGÍSTICA

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} \quad \textbf{(15.27)}$$

Como los dos valores de la variable dependiente *y* son 0 y 1, el valor de $E(y)$ en la ecuación (15.27) dará la *probabilidad* de que $y = 1$ para un conjunto dado de valores de las variables independientes $x_1, x_2, \ldots, x_p$. Como $E(y)$ se interpreta como una probabilidad, la **ecuación de regresión logística** suele expresarse de la manera siguiente.

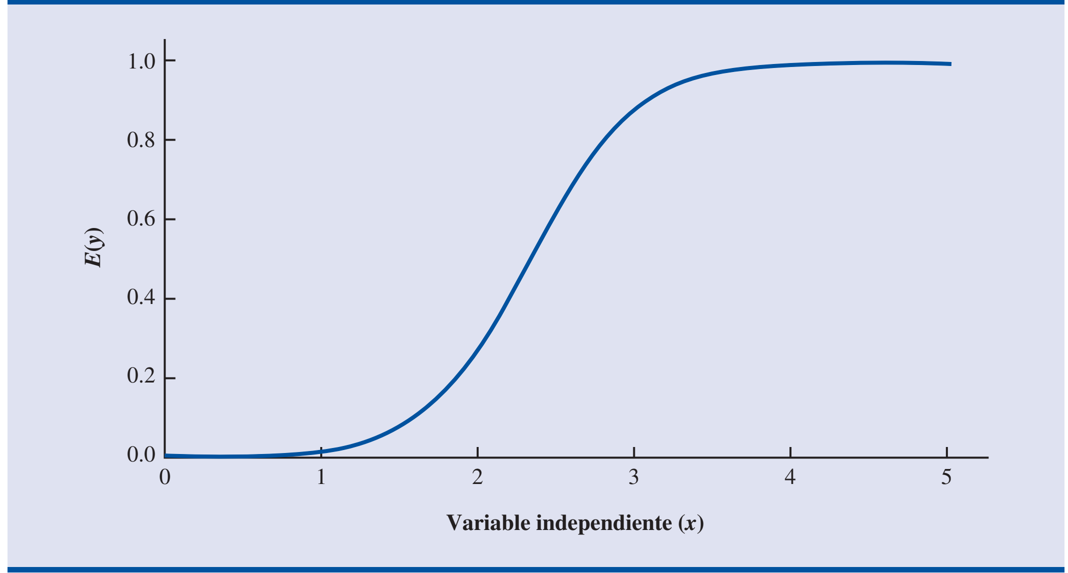
INTERPRETACIÓN DE $E(y)$ COMO PROBABILIDAD EN LA REGRESIÓN LOGÍSTICA

$$E(y) = P(y = 1 \mid x_1, x_2, \ldots, x_p) \quad \textbf{(15.28)}$$

Para entender mejor las características de la ecuación de regresión logística, supóngase que en el modelo se tiene sólo una variable independiente *x* y que los valores de los parámetros del modelo son $\beta_0 = -7$ y $\beta_1 = 3$. La ecuación de regresión logística correspondiente a estos valores de los parámetros es

$$E(y) = P(y = 1 \mid x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} = \frac{e^{-7+3x}}{1 + e^{-7+3x}} \quad \textbf{(15.29)}$$

FIGURA 15.12 ECUACIÓN DE REGRESIÓN LOGÍSTICA EN LA QUE $\beta_0 = -7$ Y $\beta_1 = 3$



En la figura 15.12 se muestra la gráfica de la ecuación (15.29). Obsérvese que la gráfica tiene forma de S. El valor de $E(y)$ va desde 0 hasta 1, aproximándose gradualmente a 1 a medida que el valor de x aumenta, y a 0 a medida que el valor de x disminuye. Obsérvese también que el valor de $E(y)$, que representa probabilidad, aumenta rápidamente al aumentar x de 2 a 3. El hecho de que los valores de $E(y)$ vayan de 0 a 1 y que la curva tenga forma de S hacen que la ecuación (15.29) sea ideal para modelar la probabilidad de que la variable dependiente sea igual a 1.

Estimación de la ecuación de regresión logística

En la regresión lineal simple y en la regresión múltiple se emplea el método de mínimos cuadrados para calcular las estimaciones $b_0, b_1, \ldots, b_p$, de los parámetros $\beta_0, \beta_1, \ldots, \beta_p$ del modelo. Debido a la forma no lineal de la ecuación de regresión logística, el método para calcular estas estimaciones es más complejo y queda fuera del alcance de este libro. Para obtener estas estimaciones se empleará un paquete de software. La **ecuación de regresión logística estimada** es

ECUACIÓN DE REGRESIÓN LOGÍSTICA ESTIMADA

$$\hat{y} = \text{estimación de } P(y = 1 \mid x_1, x_2, \ldots, x_p) = \frac{e^{b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p}}{1 + e^{b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p}} \quad \textbf{(15.30)}$$

Aquí $\hat{y}$ es una estimación de la probabilidad de que $y = 1$, para un determinado conjunto de valores de las variables independientes.

Volviendo ahora al ejemplo de Simmons Stores, las variables en este estudio están definidas como sigue:

$$y = \begin{cases} 0 \text{ si el cliente no usó el cupón} \\ 1 \text{ si el cliente usó el cupón} \end{cases}$$

$$x_1 = \text{cantidad anual gastada en Simmons Stores (en miles de \$)}$$

$$x_2 = \begin{cases} 0 \text{ si el cliente no tiene tarjeta de crédito de Simmons} \\ 1 \text{ si el cliente tiene tarjeta de crédito de Simmons} \end{cases}$$

FIGURA 15.13 RESULTADOS DE LA REGRESIÓN LOGÍSTICA PARCIAL PARA EL PROBLEMA DE LAS TIENDAS SIMMONS

En los resultados de Minitab
x_1 = Spending *(cantidad gastada)*
x_2 = Card *(tarjeta de crédito)*

```
Logistic Regression Table
                                                   Odds     95% CI
Predictor      Coef   SE Coef       Z        p    Ratio   Lower  Upper
Constant    -2.1464    0.5772   -3.72    0.000
Spending     0.3416    0.1287    2.66    0.008     1.41    1.09   1.81
Card         1.0987    0.4447    2.47    0.013     3.00    1.25   7.17

Log-Likelihood = -60.487
Test that all slopes are zero: G = 13.628, DF = 2, P-Value = 0.001
```

Por lo tanto, se elige una ecuación de regresión logística con dos variables independientes.

$$E(y) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2}}{1 + e^{\beta_0+\beta_1x_1+\beta_2x_2}} \quad \textbf{(15.31)}$$

En el apéndice 15.3 se muestra cómo se usa Minitab para generar el resultado que se muestra en la figura 15.13.

Para calcular las estimaciones de los parámetros β_0, β_1, y β_2 del modelo se aplicó el procedimiento de regresión logística binaria de Minitab a los datos muestrales de la tabla 15.11. En la figura 15.13 se muestra parte de los resultados obtenidos. Como se ve, $b_0 = -2.1464$, $b_1 = 0.3416$ y $b_2 = 1.0987$. Por lo tanto, la ecuación de regresión logística estimada es

$$\hat{y} = \frac{e^{b_0+b_1x_1+b_2x_2}}{1 + e^{b_0+b_1x_1+b_2x_2}} = \frac{e^{-2.1464+0.3416x_1+1.0987x_2}}{1 + e^{-2.1464+0.3416x_1+1.0987x_2}} \quad \textbf{(15.32)}$$

Ahora usando la ecuación (15.32) se estima la probabilidad de que un determinado tipo de clientes use el cupón. Por ejemplo, para estimar la probabilidad de que los clientes que tienen un gasto anual de \$2000 en Simmons Stores y que no tienen tarjeta de crédito de Simmons usen el cupón, en la ecuación (15.32) se sustituyen $x_1 = 2$ y $x_2 = 0$.

$$\hat{y} = \frac{e^{-2.1464+0.3416(2)+1.0987(0)}}{1 + e^{-2.1464+0.3416(2)+1.0987(0)}} = \frac{e^{-1.4632}}{1 + e^{-1.4632}} = \frac{0.2315}{1.2315} = 0.1880$$

Por lo tanto, la probabilidad estimada de que este tipo de clientes use el cupón es 0.19. De manera similar, la probabilidad de que aquellos clientes que tienen un gasto anual de \$2000 en Simmons Stores y que tiene tarjeta de crédito de Simmons usen el cupón, se estima sustituyendo en la ecuación (15.32) $x_1 = 2$ y $x_2 = 1$.

$$\hat{y} = \frac{e^{-2.1464+0.3416(2)+1.0987(1)}}{1 + e^{-2.1464+0.3416(2)+1.0987(1)}} = \frac{e^{-0.3645}}{1 + e^{-0.3645}} = \frac{0.6945}{1.6945} = 0.4099$$

Por lo tanto, la probabilidad de que los clientes de este grupo usen el cupón es aproximadamente 0.41. Parece ser que los clientes que tienen tarjeta de crédito de Simmons son los que tienen mayor probabilidad de usar el cupón. Pero antes de llegar a una conclusión, es necesario evaluar la significancia estadística de este modelo.

Prueba de significancia

La prueba de significancia en la regresión logística es similar a la prueba de significancia en la regresión múltiple. Primero se hace una prueba para probar la significancia global. En el ejemplo de Simmons Stores, las hipótesis para probar la significancia global son las siguientes:

$$H_0: \beta_1 = \beta_2 = 0$$
H_a: Uno o los dos parámetros son distintos de cero

La prueba de significancia global del modelo se basa en el valor del estadístico de prueba G. Si la hipótesis nula es verdadera, la distribución muestral de G es una distribución chi-cuadrada con grados de libertad igual al número de variables independientes en el modelo. El cálculo de G queda fuera del alcance de este libro, pero este valor de G y su correspondiente valor-p pueden obtenerse de los resultados de regresión logística binaria que da Minitab. En el último renglón de la figura 15.13 se encuentra que el valor de G es 13.628, sus grados de libertad son 2 y su correspondiente valor-p es 0.001. Por lo tanto, empleando cualquier nivel de significancia $\alpha \geq 0.001$, se rechazará la hipótesis nula y se concluirá que el modelo global es significativo.

Una vez que la prueba G ha indicado que sí existe una significancia global, para determinar si la contribución de cada una de las variables independientes al modelo es significativa, se suele realizar una prueba z. Para cada una de las variables independientes x_i las hipótesis son

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

Si la hipótesis nula es verdadera, el valor del coeficiente estimado dividido entre su error estándar seguirá una distribución de probabilidad normal estándar. En los resultados de Minitab, en la columna que tiene como título Z se presentan los valores de $z_i = b_i/s_{b_i}$ correspondientes a cada uno de los coeficientes estimados, y en la columna que tiene como título p se encuentran sus valores –p correspondientes. Supóngase que en el modelo de Simmons se emplea $\alpha = 0.05$ para probar la significancia de las variables independientes. Para la variable independiente x_1, el valor z es 2.66 y su correspondiente valor-p es 0.008. Por lo tanto, para el nivel de significancia 0.05 se rechaza H_0: $\beta_1 = 0$. De la misma manera se rechaza H_0: $\beta_2 = 0$ dado que el valor-p correspondiente a $z = 2.47$ es 0.013. Por lo tanto, empleando como nivel de significancia 0.05, ambas variables son estadísticamente significativas.

Uso en la administración

Ya se describió cómo obtener la ecuación de regresión logística estimada y cómo probar su significancia. Ahora se usará esto para hacer una recomendación para la decisión que tienen que tomar en Simmons Stores respecto a la promoción con su catálogo. Ya se calcularon $P(y = 1|x_1 = 2, x_2 = 1) = 0.4099$ y $P(y = 1|x_1 = 2, x_2 = 0) = 0.1880$. De acuerdo con estas probabilidades, se ve que entre aquellos clientes que cuyo gasto anual en Simmons Stores es de \$2000, los clientes que tienen tarjeta de crédito de Simmons tienen mayor probabilidad de hacer uso del cupón. En la tabla 15.12 se presentan las probabilidades estimadas correspondientes a clientes, tanto con tarjeta de crédito como sin tarjeta de crédito de Simmons, cuyos desembolsos anuales en Simmons Stores son desde \$1000 hasta \$7000. ¿Cómo puede Simmons emplear esta información para elegir los clientes a los que dirigirá la nueva promoción? Supóngase que Simmons desea enviar este catálogo promocional únicamente a clientes cuya probabilidad de hacer uso del cupón sea 0.40 o mayor. Empleando las probabilidades estimadas que aparecen en la tabla 15.12 la estrategia en esta promoción de Simmons sería:

Clientes que tienen tarjeta de crédito de Simmons: Enviar el catálogo a todos los clientes que durante el pasado año hayan gastado \$2000 o más.

TABLA 15.12 PROBABILIDADES ESTIMADAS PARA SIMMONS STORES

		Cantidad anual gastada						
		\$1000	**\$2000**	**\$3000**	**\$4000**	**\$5000**	**\$6000**	**\$7000**
Tarjeta de crédito	**Sí**	0.3305	0.4099	0.4943	0.5790	0.6593	0.7314	0.7931
	No	0.1413	0.1880	0.2457	0.3143	0.3921	0.4758	0.5609

Clientes sin tarjeta de crédito de Simmons: Enviar el catálogo a todos los clientes que durante el pasado año hayan gastado \$6000 o más.

Sin embargo, observando con más detenimiento las probabilidades estimadas se ve que la probabilidad de que usen el cupón aquellos clientes sin tarjeta de crédito de Simmons que gastaron \$5000 en Simmons en un año, es 0.3921. Por lo tanto, será conveniente que Simmons reconsidere su estrategia e incluya en ella a aquellos clientes que no tienen tarjeta de crédito pero que gasten en Simmons \$5000 o más en un año.

Interpretación de la ecuación de regresión logística

Para interpretar una ecuación de regresión se necesita relacionar las variables independientes de la ecuación con la cuestión de negocios a la que se trató de responder con esa ecuación. En la regresión logística, debido a que la ecuación de regresión logística no es lineal, es difícil interpretar directamente la relación entre las variables independientes y la probabilidad de que $y = 1$. Sin embargo, se ha demostrado que esta relación se puede interpretar indirectamente empleando un concepto llamado cociente de posibilidades (en inglés, odds ratio).

Las **posibilidades a favor de que ocurra un evento** se definen como la probabilidad de que ocurra el evento, dividida entre la probabilidad de que no ocurra el evento. En la regresión logística el evento de interés es siempre $y = 1$. Dado un determinado conjunto de valores de las variables independientes, las posibilidades a favor de $y = 1$ se calculan como sigue:

$$\text{odds} = \frac{P(y = 1 | x_1, x_2, \ldots, x_p)}{P(y = 0 | x_1, x_2, \ldots, x_p)} = \frac{P(y = 1 | x_1, x_2, \ldots, x_p)}{1 - P(y = 1 | x_1, x_2, \ldots, x_p)} \quad \textbf{(15.33)}$$

El **cociente de posibilidades** mide el efecto que tiene sobre estas posibilidades el aumento en una unidad de una sola de las variables independientes. El cociente de posibilidades es la probabilidad de que $y = 1$ cuando una de las variables independientes es incrementada en una unidad (odds_1) dividida entre las posibilidades de que $y = 1$ cuando no ha habido cambio en los valores de las variables independientes (odds_0).

COCIENTE DE POSIBILIDADES

$$\text{Cociente de posibilidades} = \frac{\text{odds}_1}{\text{odds}_0} \quad \textbf{(15.34)}$$

Por ejemplo, supóngase que se desean comparar las posibilidades de que use el cupón un cliente que gasta \$2000 anuales y tiene tarjeta de crédito de Simmons ($x_1 = 2$ y $x_2 = 1$) con las posibilidades de que use el cupón un cliente que gasta \$2000 anuales y no tiene tarjeta de crédito de Simmons ($x_1 = 2$ y $x_2 = 0$). Lo que interesa es interpretar el efecto que tiene un incremento de la variable independiente x_2 en una unidad. En este caso

$$\text{odds}_1 = \frac{P(y = 1 | x_1 = 2, x_2 = 1)}{1 - P(y = 1 | x_1 = 2, x_2 = 1)}$$

y

$$\text{odds}_0 = \frac{P(y = 1|x_1 = 2, x_2 = 0)}{1 - P(y = 1|x_1 = 2, x_2 = 0)}$$

Como ya se demostró, la estimación de la probabilidad de que $y = 1$ cuando $x_1 = 2$ y $x_2 = 1$ es 0.4099 y la estimación de que $y = 1$ cuando $x_1 = 2$ y $x_2 = 0$ es 0.1880. Por lo tanto,

$$\text{estimación de odds}_1 = \frac{0.4099}{1 - 0.4099} = 0.6946$$

y

$$\text{estimación de odds}_0 = \frac{0.1880}{1 - 0.1880} = 0.2315$$

La estimación del cociente de posibilidades es

$$\text{Estimación del cociente de posibilidades} = \frac{0.6946}{0.2315} = 3.00$$

Por lo tanto, se puede concluir que las posibilidades estimadas de que usen el cupón los clientes que gastaron $2000 el año pasado y tienen tarjeta de crédito de Simmons son tres veces mayores que las posibilidades estimadas de que usen el cupón los clientes que gastaron $2000 el año pasado y no tienen tarjeta de crédito de Simmons.

El cociente de posibilidades de cada una de las variables independientes se calcula manteniendo constantes todas las demás variables. Sin embargo, qué valores constantes se usen para todas las demás variables no tiene importancia. Por ejemplo, si se calcula el cociente de posibilidades para la variable tarjeta de crédito de Simmons (x_2) empleando, como valor de la variable cantidad de gasto anual (x_1), $3000 en lugar de $2000, el valor obtenido para el cociente de posibilidad estimado será el mismo (3.00). Por lo tanto, se concluye que las posibilidades estimadas de que use el cupón un cliente que tiene tarjeta de crédito de Simmons son tres veces mayores que las posibilidades estimadas de que use el cupón un cliente que no tiene tarjeta de crédito de Simmons.

El cociente de posibilidades aparece en los resultados estándar de la regresión logística proporcionados por los paquetes de software para estadística. Refiérase a los resultados de Minitab que se presentan en la figura 15.13. En la columna titulada Odds Ratio aparecen los cocientes de posibilidad estimados correspondientes a cada una de las variables independientes. El cociente de posibilidad estimada para x_1 es 1.41 y el cociente de posibilidad estimada para x_2 es 3.00. Ya se indicó antes cómo interpretar el coeficiente de posibilidad estimada en el caso de la variable binaria independiente x_2. Ahora se considerará la interpretación del cociente de posibilidades en el caso de la variable continua independiente x_1.

El valor 1.41 en la columna titulada Odds Ratio (cociente de posibilidades) de los resultados de Minitab indica que la posibilidad estimada de que use el cupón un cliente que gastó $3000 durante el año pasado es 1.41 veces mayor que la probabilidad estimada de que use el cupón un cliente que gastó $2000 durante el año pasado. Más aún, esta interpretación es correcta para cualquier cambio en una unidad de x_1. Por ejemplo, las posibilidades estimadas de que use el cupón un cliente cuyo gasto anual el año pasado fue $5000 son 1.41 veces mayores que las probabilidades estimadas de que use el cupón un cliente cuyo gasto anual el año pasado fue $4000. Pero, supóngase que interesa la variación en las posibilidades cuando hay un incremento de más de una unidad en cualquiera de las variables independientes. Obsérvese que x_1 toma valores desde 1 hasta 7. El cociente de posibilidades presentado en los resultados de Minitab no responde a esta pregunta. Para responder a esta pregunta es necesario explorar la relación entre el cociente de posibilidades y los coeficientes de regresión.

Existe una relación única entre el cociente de posibilidades de una variable y su correspondiente coeficiente de regresión. Se puede demostrar que para toda variable independiente de una ecuación de regresión logística

$$\text{Cociente de posibilidades} = e^{\beta_i}$$

Para ilustrar esta relación, empleando el ejemplo de Simmons Stores considérese la variable independiente x_1. El cociente de posibilidades estimado para x_1 es

$$\text{Cociente de posibilidades estimado} = e^{b_1} = e^{0.3416} = 1.41$$

De manera similar, el cociente de posibilidades estimado para x_2 es

$$\text{Cociente de posibilidades estimado} = e^{b_2} = e^{1.0987} = 3.00$$

Esta relación entre el cociente de posibilidades y los coeficientes de las variables independientes facilitan el cálculo del cociente de posibilidades una vez obtenidas las estimaciones de los parámetros del modelo. Además, esto permite también investigar cambios en el cociente de posibilidades cuando hay cambios mayores o menores a una unidad en una de las variables continuas independientes.

El cociente de posibilidades de una variable independiente representa la variación que hay en las posibilidades por una variación de una unidad en la variable independiente permaneciendo constantes todas las demás variables independientes. Supóngase que se desea conocer el efecto de una variación de más de una unidad, por ejemplo de c unidades. Supóngase que, en el ejemplo de Simmons, se quieren comparar las posibilidades de que use el cupón un cliente que gasta \$5000 anuales ($x_1 = 5$) con las posibilidades de que use el cupón un cliente que gasta \$2000 anuales ($x_1 = 2$). En este caso $c = 5 - 2 = 3$ y el correspondiente cociente de posibilidades es

$$e^{cb_1} = e^{3(0.3416)} = e^{1.0248} = 2.79$$

Esto indica que las posibilidades estimadas de que usen el cupón los clientes cuyo gasto anual es de \$5000 son 2.70 veces mayores que las posibilidades estimadas de que usen el cupón los clientes cuyo gasto anual es de \$2000. En otras palabras, el cociente de posibilidades estimado para un aumento de \$3000 en los gastos anuales es 2.79.

En general, el cociente de posibilidades permite comparar las posibilidades de dos eventos diferentes. Si el cociente de posibilidades es 1, los dos eventos tienen las mismas posibilidades. Por lo tanto, si la variable independiente que se considera (como, por ejemplo, el estatus respecto a la tarjeta de crédito de Simmons) tiene efecto positivo sobre la probabilidad de que ocurra el evento el cociente de posibilidades correspondiente será mayor a 1. La mayor parte de los paquetes de software para estadística dan también un intervalo de confianza para el cociente de posibilidades. En la figura 15.13 los resultados de Minitab presentados proporcionan un intervalo de 95% de confianza para cada uno de los cocientes de posibilidades. Por ejemplo, la estimación puntual del cociente de posibilidad de x_1 es 1.41 y el intervalo de 95% de confianza va de 1.09 a 1.81.Como el intervalo de confianza no contiene el valor 1, se concluye que x_1 tiene un efecto significativo sobre el cociente de posibilidades estimado. De manera similar, el intervalo de 95% de confianza para el cociente de posibilidades de x_2 va de 1.25 a 7.17. Como este intervalo tampoco contiene el valor 1, se puede concluir también que x_2 tiene un efecto significativo sobre el cociente de posibilidades.

Transformación logit

Entre las posibilidades a favor de $y = 1$ y el exponente de e en la ecuación de regresión logística, se puede observar una interesante relación. Se puede demostrar que

$$\ln(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Esta ecuación indica que el logaritmo natural de las posibilidades a favor de $y = 1$ es función lineal de las variables independientes. A esta función lineal se le llama **logit**. Para denotar el logit se empleará la notación $g(x_1, x_2, \ldots, x_p,)$.

LOGIT

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad \textbf{(15.35)}$$

Sustituyendo en la ecuación (15.27) $\beta_1 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ por $g(x_1, x_2, \ldots, x_p,)$, la ecuación de regresión logística se puede expresar como

$$E(y) = \frac{e^{g(x_1, x_2, \ldots, x_p)}}{1 + e^{g(x_1, x_2, \ldots, x_p)}} \tag{15.36}$$

Una vez estimados los parámetros de la ecuación de regresión logística, se puede calcular una estimación del logit. Empleando $\hat{g}(x_1, x_2, \ldots, x_p)$ para denotar el **logit estimado**, se tiene

LOGIT ESTIMADO

$$\hat{g}(x_1, x_2, \ldots, x_p) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \tag{15.37}$$

Por lo tanto, en términos del logit estimado, la ecuación de regresión estimada es

$$\hat{y} = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}} = \frac{e^{\hat{g}(x_1, x_2, \ldots, x_p)}}{1 + e^{\hat{g}(x_1, x_2, \ldots, x_p)}} \tag{15.38}$$

En el ejemplo de Simmons Stores, el logit estimado es

$$\hat{g}(x_1, x_2) = -2.1464 + 0.3416x_1 + 1.0987x_2$$

y la ecuación de regresión estimada es

$$\hat{y} = \frac{e^{\hat{g}(x_1, x_2)}}{1 + e^{\hat{g}(x_1, x_2)}} = \frac{e^{-2.1464 + 0.3416x_1 + 1.0987x_2}}{1 + e^{-2.1464 + 0.3416x_1 + 1.0987x_2}}$$

Por lo tanto, debido a la relación única que existe entre el logit estimado y la ecuación de regresión logística estimada, es posible calcular las probabilidades estimadas para Simmons Stores dividiendo $e^{\hat{g}(x_1, x_2)}$ entre $1 + e^{\hat{g}(x_1, x_2)}$.

NOTAS Y COMENTARIOS

1. Debido a la relación única que existe entre los coeficientes estimados del modelo y los correspondientes cocientes de posibilidades, la prueba general de significancia basada en el estadístico G es también una prueba general de significancia para los cocientes de posibilidades. Además, la prueba z para la significancia de cada uno de los parámetros del modelo es también una prueba estadística para los correspondientes cocientes de posibilidades.
2. En la regresión simple y en la regresión múltiple, se usa el coeficiente de determinación para medir la bondad de ajuste. En la regresión logística no hay una sola medida que tenga una interpretación similar. El estudio de la bondad de ajuste queda fuera del alcance de esta introducción a la regresión logística.

Ejercicios

Aplicaciones



44. Vaya al ejemplo de Simmons Stores presentado en esta sección. La variable dependiente es $y = 1$ si el cliente usó el cupón y $y = 0$ si no lo usó. Supóngase que la única información de que se dispone para predecir si un cliente usará o no el cupón es el estatus del cliente respecto a la posesión de una tarjeta de crédito de la empresa, que es $x = 1$ si el cliente tiene tarjeta de crédito de Simmons y $x = 0$ si no es así.
    a. Dé la ecuación de regresión logística que relaciona x y y.
    b. ¿Cuál es la interpretación de $E(y)$ cuando $x = 0$?

c. Empleando los datos de Simmons presentados en la tabla 15.11, use Minitab para calcular el logit estimado.
d. Empleando el logit estimado obtenido en el inciso c), obtenga una estimación de la probabilidad de que usen el cupón los clientes que no tienen tarjeta de crédito de Simmons y una estimación de la probabilidad de que usen el cupón los clientes que no tienen tarjeta de crédito de Simmons.
e. De la estimación del cociente de posibilidades. ¿Cuál es su interpretación?

45. En la tabla 15.12 se proporcionaron estimaciones de las probabilidades de uso del cupón en la promoción mediante catálogo de Simmons Stores. Para cada combinación de valores de las variables independientes se obtuvo un valor diferente.
a. Calcule las posibilidades de que use el cupón un cliente cuyo gasto anual en Simmons es de $4000 y que no tiene tarjeta de crédito de Simmons ($x_1 = 4$, $x_2 = 0$).
b. Use la información de la tabla 15.12 y el inciso a) para calcular el cociente de posibilidades para la variable tarjeta de crédito de Simmons $x_2 = 0$ manteniendo constantes los gastos anuales en $x_1 = 4$.
c. En el libro, el cociente de posibilidades para la variable tarjeta de crédito se calculó empleando la información presentada en la columna $2000 de la tabla 15.12. ¿Obtuvo, en el inciso b), la misma información para el valor del cociente de posibilidades?

46. El Community Bank desea aumentar la cantidad de clientes a los que les depositan directamente su nómina. El gerente está considerando una campaña que requerirá que cada gerente de sucursal llame a cada cliente que no reciba directamente su nómina. Como incentivo para aceptar recibir directamente su nómina, se les ofrecerá revisión gratuita de su cuenta durante dos años. Debido al tiempo y a los costos de esta campaña, el gerente desea que esta campaña se dirija a aquellos clientes que tengan la mayor probabilidad de aceptar recibir directamente su nómina. El gerente piensa que el saldo promedio mensual en la cuenta de cheques del cliente puede ser un predictor útil para determinar si un cliente aceptará o no recibir directamente su nómina. Para investigar la relación entre estas dos variables, Community Bank prueba la nueva campaña utilizando una muestra de cuentas de cheques de 50 clientes que actualmente no reciben directamente su nómina. En los datos muestrales se presenta el saldo mensual promedio en la cuenta de cheques (en miles de dólares) y si el cliente aceptó recibir directamente el depósito de su nómina (1 significa aceptó el depósito directo de su nómina y 0 significa el cliente no aceptó el depósito directo de su nómina). Estos datos se encuentran en el archivo Bank del disco compacto; a continuación se presenta parte de estos datos.

archivo en CD
Bank

Cliente	x = saldo mensual	y = depósito directo
1	1.22	0
2	1.56	0
3	2.10	0
4	2.25	0
5	2.89	0
6	3.55	0
7	3.56	0
8	3.65	1
.	.	.
.	.	.
.	.	.
48	18.45	1
49	24.98	0
50	26.05	1

a. Dé la ecuación de regresión logística que relaciona x y y.
b. Empleando los datos de Community Bank, use Minitab para calcular la ecuación de regresión logística estimada.
c. Realice una prueba de significancia empleando el estadístico de prueba G. Use $\alpha = 0.05$.

d. Estime la probabilidad de que los clientes cuyo saldo mensual promedio sea $1000 acepten recibir directamente el depósito de su nómina.
e. Supóngase que Community Bank desea contactar únicamente a los clientes para los que la probabilidad de aceptar recibir directamente su nómina sea de 0.50 o mayor. ¿Cuál es el saldo promedio requerido para tener esta probabilidad?
f. Dé la estimación del cociente de posibilidades. ¿Cuál es su interpretación?

47. En los últimos años en Lakeland Collage ha aumentado el porcentaje de estudiantes que abandonan sus estudios después del primer año. El año pasado, Lakeland Collage inició un programa voluntario de orientación para ayudar a los estudiantes de primer año a que se adapten a la vida del campus. Si Lakeland Collage demuestra que ese programa tiene resultados positivos, se considerará la posibilidad de que el programa sea obligatorio para todos los estudiantes de primer año. La administración de Lakeland Collage supone que los estudiantes que tienen GPA bajo son los que tienen mayor probabilidad de abandonar los estudios al final del primer año. Con objeto de investigar la relación de estas variables con la permanencia de los estudiantes en la escuela, Lakeland Collage tomó una muestra aleatoria de 100 estudiantes de primer año. Los datos se encuentran en el archivo Lakeland del disco compacto; a continuación se reproduce parte de esos datos.



Lakeland

Estudiante	GPA	Programa	Resultado
1	3.78	1	1
2	2.38	0	1
3	1.30	0	0
4	2.19	1	0
5	3.22	1	1
6	2.68	1	1
.	.	.	.
.	.	.	.
.	.	.	.
98	2.57	1	1
99	1.70	1	1
100	3.85	1	1

La variable dependiente toma el valor $y = 1$ si el estudiante permanece en la escuela y $y = 0$ si no es así. Las dos variables independientes son

$$x_1 = \text{GPA al final del primer semestre}$$

$$x_2 = \begin{cases} 0 \text{ si el estudiante participa en el programa de orientación} \\ 1 \text{ si el estudiante no participa en el programa de orientación} \end{cases}$$

a. Dé la ecuación de regresión logística que relaciona x_1 y x_2 con y.
b. Dé la interpretación de $E(y)$ cuando $x_2 = 0$.
c. Use las dos variables independientes y Minitab para calcular el logit estimado.
d. Realice una prueba de significancia global empleando $\alpha = 0.05$.
e. Empleando $\alpha = 0.05$, determine si cada una de las variables independientes son significativas.
f. Use el logit estimado del inciso c) para obtener una estimación de la probabilidad de que un estudiante cuyo GPA es 2.5 y que no participó en el programa de orientación permanezca en la escuela. ¿Cuál es la estimación de esta probabilidad para un estudiante cuyo GPA es 2.5 y que sí participó en el programa de orientación?
g. Dé la estimación del cociente de posibilidades para el programa de orientación. Interprételo.
h. ¿Recomendaría convertir el programa de orientación en un programa obligatorio? ¿Por qué sí o por qué no?

48. *Consumer Report* le realizó una prueba de sabor a 19 marcas de chocolates. En los datos a continuación se da el precio por porción, en base al tamaño de porción de la FDA que es de 1.4 onzas, así como una evaluación de la calidad de los 19 chocolates tomados para la prueba (*Consumer Report,* febrero 2002).



Fabricante	Precio	Evaluación
Bernard Callebaut	3.17	muy bueno
Candinas	3.58	excelente
Fannie May	1.49	bueno
Godiva	2.91	muy bueno
Hershey's	0.76	bueno
L.A. Burdick	3.70	muy bueno
La Maison du Chocolate	5.08	excelente
Leonidas	2.11	muy bueno
Lindt	2.20	bueno
Martine's	4.76	excelente
Michael Recchiuti	7.05	muy bueno
Neuchatel	3.36	bueno
Neuchatel Sugar Free	3.22	bueno
Richard Donnelly	6.55	muy bueno
Russell Stover	0.70	bueno
See's	1.06	muy bueno
Teuscher Lake of Zurich	4.66	muy bueno
Whitman's	0.70	regular
Whitman's Sugar Free	1.21	regular

Suponga que desea determinar si los productos que son más caros son mejor evaluados. Para los propósitos de este ejercicio, emplee la siguiente variable binaria dependiente.

$y = 1$ si la evaluación de la calidad fue excelente o muy buena y
$y = 0$ si la evaluación de la calidad fue buena o regular.

a. Dé la ecuación de regresión logística que relaciona x = precio por porción con y.
b. Use Minitab para calcular el logit estimado.
c. Use el logit estimado que obtuvo en el inciso b) para obtener una estimación de la probabilidad de que la evaluación de un chocolate cuyo precio por porción es $4.00 sea muy bueno o excelente.
d. Dé la estimación del cociente de posibilidades. Dé su interpretación.

Resumen

En este capítulo se presentó la regresión múltiple como extensión del análisis de regresión lineal simple presentado en el capítulo 14. El análisis de regresión múltiple permite entender cómo está relacionada una variable dependiente con dos o más variables independientes. La ecuación de regresión $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ indica que el valor esperado o valor medio de la variable dependiente y está relacionado con los valores de las variables independientes $x_1, x_2, \ldots, x_p$. Para obtener la ecuación de regresión estimada $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$ se emplean los datos muestrales y el método de mínimos cuadrados. En efecto $b_1, b_2, \ldots, b_p$ son estadísticos muestrales que se emplean para estimar los parámetros desconocidos $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ del modelo. Para hacer énfasis en el hecho de que los paquetes de software para estadística son los únicos medios realistas para realizar los numerosos cálculos que se requieren en el análisis de regresión múltiple, a lo largo de todo el capítulo se emplearon las impresiones de computadora.

El coeficiente de determinación múltiple se presentó como una medida de la bondad de ajuste de la ecuación de regresión estimada. Este coeficiente determina la proporción de la variación en y que puede ser explicada por la ecuación de regresión estimada. El coeficiente de determinación múltiple ajustado es una medida similar de bondad de ajuste que se adapta al número de variables independientes evitando, de esta manera, sobreestimar el efecto de la adición de más variables independientes.

Como un medio para determinar estadísticamente si la relación entre las variables era significativa se presentaron una prueba F y una prueba t. La prueba F sirve para determinar si existe

una relación global significativa entre la variable dependiente y el conjunto de todas las variables independientes. La prueba t se usa para determinar si existe una relación significativa entre la variable dependiente y una determinada variable independiente del modelo de regresión. Se trató la relación entre las variables independientes, a lo cual se le llama multicolinearidad.

En la sección sobre variables cualitativas independientes se mostró el uso de variables ficticias para incorporar datos cualitativos en el análisis de regresión múltiple. En la sección sobre análisis residual se mostró el uso del análisis residual para confirmar las suposiciones del modelo, detectar observaciones atípicas e identificar observaciones influyentes. Se estudiaron los residuales estandarizados, la influencia, los residuales eliminados estudentizados y la medida de la distancia de Cook. El capítulo se concluyó con una sección sobre el uso de la regresión logística para modelar situaciones en las que la variable dependiente sólo puede asumir dos valores.

Glosario

Análisis de regresión múltiple Análisis de regresión en el que hay dos o más variables independientes.

Modelo de regresión múltiple Ecuación matemática que describe cómo está relacionada la variable dependiente y con las variables independientes $x_1, x_2, \ldots, x_p$ y con el término del error ϵ.

Ecuación de regresión múltiple Ecuación matemática que relaciona el valor esperado o valor medio de la variable dependiente con los valores de las variables independientes. Es decir $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.

Ecuación de regresión múltiple estimada Estimación de la ecuación de regresión múltiple que se basa en datos muestrales y en el método de mínimos cuadrados; es $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$.

Método de mínimos cuadrados Método empleado para obtener la ecuación de regresión estimada. Mediante este método se minimiza la suma de los cuadrados de los residuales (la desviación (diferencia) que existe entre los valores observados de la variable dependiente, y_i, y los valores estimados de la variable dependiente $\hat{y}_i$)

Coeficiente de determinación múltiple Medida de la bondad de ajuste de la ecuación de regresión múltiple estimada. Se puede interpretar como la proporción en la variabilidad de la variable dependiente que es explicada por la ecuación de regresión estimada.

Coeficiente de determinación múltiple ajustado Medida de la bondad de ajuste de la ecuación de regresión múltiple estimada, pero modifica de acuerdo con el número de variables independientes en el modelo para evitar, de esta manera, sobreestimar el efecto que tiene agregar más variables independientes.

Multicolinealidad Término usado para describir la correlación entre las variables independientes.

Variable cualitativa independiente Variable independiente con datos cualitativos.

Variable ficticia Variable usada para modelar el efecto de las variables cualitativas independientes. Las variables ficticias sólo toman los valores cero y uno.

Influencia Mide qué tan lejos de su media se encuentran los valores de las variables independientes.

Observación atípica Observación que se sale del patrón que sigue el resto de las observaciones.

Residuales eliminados estudentizados Residuales estandarizados que se basan en un error estándar de estimación corregido que se obtuvo al eliminar la observación i del conjunto de datos y realizar después el análisis de regresión y los cálculos.

Observación influyente Observación que tiene una gran influencia en los resultados de la regresión.

Medida de la distancia de Cook Medida de la influencia de una observación que se basa tanto en la influencia (leverage) de la observación i como en el residual de la observación i.

Ecuación de regresión logística Ecuación matemática que relaciona $E(y)$, la probabilidad de que $y = 1$, con los valores de las variables independientes; es decir $E(y) = P(y = 1 | x_1, x_2, \ldots, x_p) = \dfrac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}$.

Ecuación de regresión logística estimada Estimación de la ecuación de regresión logística que se basa en datos muestrales; es decir $\hat{y}$ = estimación de $P(y = 1|x_1, x_2, \ldots, x_p) = \dfrac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}{1 + e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}$.

Posibilidades a favor de la ocurrencia de un evento Probabilidad de que ocurra el evento entre la probabilidad de que no ocurra el evento.

Cociente de posibilidades El cociente que se obtiene al dividir la posibilidad de que $y = 1$ dado que una de las variables independientes aumentó en una unidad (odds_1) entre la posibilidad de que $y = 1$ dado que no hay ninguna variación en los valores de las variables independientes (odds_0); es decir, cociente de posibilidades (odds ratio) = $(\text{odds}_1)/(\text{odds}_0)$.

Logit Logaritmo natural de las posibilidades a favor de $y = 1$; es decir $g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p$.

Logit estimado Estimación del logit basada en datos muestrales; es decir $\hat{g}(x_1, x_2, \ldots, x_p) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$.

Fórmulas clave

Modelo de regresión múltiple

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p + \epsilon \tag{15.1}$$

Ecuación de regresión múltiple

$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p \tag{15.2}$$

Ecuación de regresión múltiple estimada

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p \tag{15.3}$$

Criterio de mínimos cuadrados

$$\min \Sigma(y_i - \hat{y}_i)^2 \tag{15.4}$$

Relación entre STC, SCR y SCE

$$\text{STC} = \text{SCR} + \text{SCE} \tag{15.7}$$

Coeficiente de determinación múltiple

$$R^2 = \frac{\text{SCR}}{\text{STC}} \tag{15.8}$$

Coeficiente de determinación múltiple ajustado

$$R_a^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1} \tag{15.9}$$

Cuadrado medio debido a la regresión

$$\text{CMR} = \frac{\text{SCR}}{p} \tag{15.12}$$

Cuadrado medio debido al error

$$\text{CME} = \frac{\text{SCE}}{n - p - 1} \tag{15.13}$$

Estadístico de prueba *F*

$$F = \frac{\text{CMR}}{\text{CME}} \tag{15.14}$$

Estadístico de prueba *t*

$$t = \frac{b_i}{s_{b_i}} \tag{15.15}$$

Residual estandarizado de la observación *i*

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{15.23}$$

Desviación estándar del residual *i*

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \tag{15.24}$$

Medida de la distancia de Cook

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(p - 1)s^2}\left[\frac{h_i}{(1 - h_i)^2}\right] \tag{15.25}$$

Ecuación de regresión logística

$$E(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}} \tag{15.27}$$

Ecuación de regresión logística estimada

$$\hat{y} = \text{estimación de } P(y = 1 \mid x_1, x_2, \ldots, x_p) = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}} \tag{15.30}$$

Cociente de posibilidades (*odds ratio*)

$$\text{Cociente de posibilidades } (\textit{Odds ratio}) = \frac{\text{odds}_1}{\text{odds}_0} \tag{15.34}$$

Logit

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{15.35}$$

Logit estimado

$$\hat{g}(x_1, x_2, \ldots, x_p) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \tag{15.37}$$

Ejercicios complementarios

49. El departamento de admisión de Clearwater Collage obtuvo la siguiente ecuación de regresión estimada en la que relaciona el promedio final obtenido en la universidad (GPA) con la puntuación del estudiante en el área de matemáticas del examen de admisión a la universidad (SAT) y con su promedio final (GAP) en bachillerato.

$$\hat{y} = -1.41 + 0.0235x_1 + 0.00486x_2$$

donde

x_1 = promedio final en el bachillerato
x_2 = puntuación en el área de matemáticas admisión (SAT)
y = promedio final en la universidad

a. Interprete los coeficientes de esta ecuación de regresión estimada.
b. Estime el promedio final (GPA) en la universidad de un estudiante cuyo promedio en el bachillerato sea 84 y cuya puntuación en el área de matemáticas del examen de admisión (SAT) es 540.

50. El director de personal de Electronics Associates obtuvo la siguiente ecuación de regresión estimada que relaciona la puntuación obtenida por un empleado en un examen sobre su satisfacción con el trabajo con la antigüedad y el salario del empleado.

$$\hat{y} = 14.4 - 8.69x_1 + 13.5x_2$$

donde

x_1 = antigüedad (años)
x_2 = salario (dólares)
y = puntuación en el examen sobre satisfacción con el trabajo (puntuaciones más altas corresponden a mayor satisfacción con el trabajo)

a. Interprete los coeficientes de esta ecuación de regresión estimada.
b. Estime la puntuación obtenida en el examen sobre satisfacción con el trabajo que tendrá un empleado cuya antigüedad es de cuatro años y que gana $6.50 por hora.

51. A continuación se presentan los resultados, incompletos, obtenidos con un paquete de software para un análisis de regresión.

```
The regression equation is
Y = 8.103 + 7.602 X1 + 3.111 X2

Predictor          Coef       SE Coef        T
Constant         _______        2.667      _____
X1               _______        2.105      _____
X2               _______        0.613      _____

S = 3.335      R-sq = 92.3%     R-sq(adj) = _____%

Analysis of Variance

SOURCE            DF          SS         MS         F
Regression      ______       1612      ______     _____
Residual Error     12      ______      ______
Total           ______     ______
```

a. Calcule los cocientes t adecuados
b. Pruebe la significancia de β_1 y β_2 empleando $\alpha = 0.05$.
c. Calcule las cantidades que faltan en las columnas DF, SS y MS.
d. Calcule R^2_a.

52. En el ejercicio 49, se vio que el departamento de admisión de Clearwater Collage obtuvo la siguiente ecuación de regresión estimada en la que relacionaba el promedio final obtenido en la universidad (GPA) con la puntuación de un estudiante en el área de matemáticas del examen de admisión a la universidad (SAT) y con su promedio final (GAP) en bachillerato.

$$\hat{y} = -1.41 + 0.0235x_1 + 0.00486x_2$$

donde

x_1 = promedio final en el bachillerato
x_2 = puntuación en el área de matemáticas del examen de admisión (SAT)
y = promedio final en la universidad

A continuación se presentan los resultados, incompletos, obtenidos con Minitab.

```
The regression equation is
Y = -1.41 + .0235 X1 + .00486 X2

Predictor           Coef       SE Coef          T
Constant         -1.4053        0.4848      _____
X1              0.023467      0.008666      _____
X2               _______      0.001077      _____

S = 0.1298      R-sq = ______     R-sq(adj) = ______

Analysis of Variance

SOURCE            DF         SS         MS        F
Regression     _____    1.76209      _____    _____
Residual Error _____    _______      _____
Total              9    1.88000
```

a. Calcule las cantidades faltantes en estos resultados.
b. Calcule F y empleando como nivel de significancia $a = 0.05$ pruebe si existe una relación significativa.
c. ¿Proporciona la ecuación de regresión estimada un buen ajuste a los datos? Explique.
d. Use la prueba t y $\alpha = 0.05$ para probar H_0: $\beta_1 = 0$ y H_0: $\beta_2 = 0$.

53. En el ejercicio 50 se vio que el director de personal de Electronics Associates obtuvo la siguiente ecuación de regresión estimada que relacionaba la puntuación obtenida por un empleado en un examen sobre su satisfacción con el trabajo con la antigüedad y el salario del empleado.

$$\hat{y} = 14.4 - 8.69x_1 + 13.5x_2$$

donde

x_1 = antigüedad (años)
x_2 = salario (dólares)
y = puntuación en el examen sobre satisfacción con el trabajo (puntuaciones más altas corresponden a mayor satisfacción con el trabajo)

A continuación se presentan los resultados, incompletos, obtenidos con Minitab

```
The regression equation is
Y = 14.4 – 8.69 X1 + 13.52 X2

Predictor            Coef        SE Coef          T
Constant           14.448          8.191       1.76
X1                _______          1.555     ______
X2                 13.517          2.085     ______

S = 3.773      R–sq = ______%  R–sq(adj) = ______%

Analysis of Variance

SOURCE              DF           SS         MS         F
Regression           2      _______     ______    ______
Residual Error  ______        71.17     ______
Total                7        720.0
```

a. Calcule las cantidades faltantes en estos resultados.
b. Calcule F y empleando como nivel de significancia 0.05 pruebe si la relación es significativa.
c. ¿Proporciona la ecuación de regresión estimada un buen ajuste a los datos? Explique.
d. Use la prueba t y $\alpha = 0.05$ para probar H_0: $\beta_1 = 0$ y H_0: $\beta_2 = 0$.

54. La revista *SmartMoney* evaluó 65 zonas metropolitanas para determinar si el valor de las casas (*home values*) estaba cambiando (are headed). La puntuación para una ciudad ideal era 100 y significaba que todos los factores medidos eran tan favorables como era posible. Zonas cuya puntuación era 60 o más, eran zonas en las que era posible una revalorización de los precios; zonas cuya puntuación era menor a 50 eran zonas que podrían ver una disminución en el valor de la vivienda. Dos de los factores evaluados fueron resistencia a la recesión y accesibilidad de la zona. Estos dos factores se evaluaron empleando una escala de 0 (evaluación más baja) a 10 (evaluación más alta). A continuación se presentan los datos obtenidos en una muestra de 20 ciudades evaluadas por *SmartMoney* (*SmartMoney*, febrero de 2002).

archivo en CD
HomeValue

Área metropolitana	Resistencia a la recesión	Accesibilidad	Puntuación
Tucson	10	7	70.7
Fort Worth	10	7	68.5
San Antonio	6	8	65.5
Richmond	8	6	63.6
Indianapolis	4	8	62.5
Philadelphia	0	10	61.9
Atlanta	2	6	60.7
Phoenix	4	5	60.3
Cincinnati	2	7	57.0
Miami	6	5	56.5
Hartford	0	7	56.2
Birmingham	0	8	55.7
San Diego	8	2	54.6
Raleigh	2	7	50.9
Oklahoma City	1	6	49.6
Orange County	4	2	49.1
Denver	4	4	48.6
Los Ángeles	0	7	45.7
Detroit	0	5	44.3
Nueva Orleáns	0	5	41.2

a. Dé una ecuación de regresión estimada que sirva para dar la puntuación conociendo la resistencia a la recesión. Empleando como nivel de significancia 0.05, pruebe la significancia de la relación
b. ¿Proporciona la ecuación obtenida en el inciso a) un buen ajuste a los datos? Explique.
c. Obtenga una ecuación de regresión estimada que sirva para predecir la puntuación a partir de la resistencia a la recesión y la accesibilidad. Empleando como nivel de significancia 0.05 pruebe la significancia global.

55. *Consumer Reports* examinó ampliamente y presentó las evaluaciones de 24 caminadoras. A cada caminadora se le dio una calificación general que se basaba principalmente en su facilidad de uso, ergonomía, gama de ejercicio y calidad. En general, una mejor calificación corresponde a un mejor funcionamiento. En la información a continuación se presenta el precio, la evaluación de la calidad y la puntuación general de las 24 caminadoras (*Consumer Reports*, febrero de 2006).



Treadmills

Marca y modelo	Precio	Calidad	Calificación
Landice L7	2900	Excelente	86
NordicTrack S3000	3500	Muy buena	85
SportsArt 3110	2900	Excelente	82
Precor	3500	Excelente	81
True Z4 HRC	2300	Excelente	81
Vision Fitness T9500	2000	Excelente	81
Precor M 9.31	3000	Excelente	79
Vision Fitness T9200	1300	Muy buena	78
Star Trac TR901	3200	Muy buena	72
Trimline T350HR	1600	Muy buena	72
Schwinn 820p	1300	Muy buena	69
Bowflex 7-Series	1500	Excelente	83
NordicTrack S1900	2600	Muy buena	83
Horizon Fitness PST8	1600	Muy buena	82
Horizon Fitness 5.2T	1800	Muy buena	80
Evo by Smooth Fitness FX30	1700	Muy buena	75
ProForm 1000S	1600	Muy buena	75
Horizon Fitness CST4.5	1000	Muy buena	74
Keys Fitness 320t	1200	Muy buena	73
Smooth Fitness 7.1HR Pro	1600	Muy buena	73
NordicTrack C2300	1000	Bueno	70
Spirit Inspire	1400	Muy buena	70
ProForm 750	1000	Buena	67
Image 19.0 R	600	Buena	66

a. Con estos datos obtenga una ecuación de regresión estimada que sirva para estimar la calificación general cuando se conoce el precio.
b. Use $\alpha = 0.05$ para probar la significancia general.
c. Para incorporar el efecto de la calidad, una variable cualitativa de tres niveles, se emplearon dos variables ficticias. Calidad-E y Calidad-MB. Cada variable toma los valores 0 y 1 como sigue.

$$\text{Calidad-E} = \begin{cases} 1 \text{ si evaluación de la calidad es excelente} \\ 0 \text{ si no es así} \end{cases}$$

$$\text{Calidad-MB} = \begin{cases} 1 \text{ si evaluación de la calidad es muy buena} \\ 0 \text{ si no es así} \end{cases}$$

Obtenga una ecuación de regresión estimada que sirva para estimar la puntuación general cuando se conoce el precio y la evaluación de la calidad.

d. Empleando $\alpha = 0.10$ pruebe la significancia general de la ecuación de regresión estimada obtenida en el inciso c).
e. Use la prueba *t* para determinar la significancia de cada una de las variables independientes de la ecuación de regresión estimada obtenida en el inciso c). Use $\alpha = 0.10$.
f. Dé la gráfica de los residuales estandarizados. ¿Parece razonable la forma de la gráfica de residuales?
g. ¿Hay en estos datos alguna observación atípica o alguna observación influyente?
h. Estime la calificación general dada a una caminadora cuyo precio es $2000 y que como evaluación de su calidad obtuvo buena. ¿Cuánto varía esta estimación si la evaluación de la calidad es muy buena? Explique.

56. La *Fuel Economy Guide* (guía de economía de combustible) del Departamento de energía de Estados Unidos proporciona datos sobre el rendimiento del combustible en automóviles y camiones. A continuación se presenta una parte de los datos obtenidos para 35 camiones furgonetas estándar producidos por Chevrolet y General Motors (www.fueleconomy.gov, 21 de marzo de 2003). En la columna titulada tracción se indica si el vehículo tiene tracción en dos ruedas (2WD) o tracción en cuatro ruedas (4WD). En la columna titulada desplazamiento se da el desplazamiento del motor en litros, en la columna titulada cilindros se especifica la cantidad de cilindros que tiene el motor, y en la columna titulada transmisión se indica si la furgoneta tiene transmisión automática o manual. En la columna titulada ciudad mpg se da el rendimiento de combustible en ciudad, en millas por galón (mpg).

archivo en CD
FuelEcon

Camión	Nombre	Tracción	Desplazamiento	Cilindros	Transmisión	Ciudad MPG
1	C1500 Silverado	2WD	4.3	6	Auto	15
2	C1500 Silverado	2WD	4.3	6	Manual	15
3	C1500 Silverado	2WD	4.8	8	Auto	15
4	C1500 Silverado	2WD	4.8	8	Manual	16
5	C1500 Silverado	2WD	5.3	8	Auto	11
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
32	K1500 Sierra	4WD	5.3	8	Auto	15
33	K1500 Sierra	4WD	5.3	8	Auto	15
34	Sonoma	4WD	4.3	6	Auto	17
35	Sonoma	4WD	4.3	6	Manual	15

a. Dé una ecuación de regresión estimada que sirva para predecir el consumo de combustible en la ciudad cuando se conoce el desplazamiento del motor. Haga una prueba de significancia empleando $\alpha = 0.05$.
b. Agregue la variable ficticia Tracción4, que toma el valor 0 si la furgoneta tiene tracción en dos ruedas y 1 si tiene tracción en cuatro ruedas. Obtenga una ecuación de regresión estimada que sirva para predecir el consumo de combustible en ciudad cuando se conoce el desplazamiento del motor y el valor de la variable ficticia Tracción4.
c. Use $\alpha = 0.05$ para determinar si la variable ficticia agregada en el inciso b) es significativa.
d. Agregue la variable ficticia OchoCil, que toma el valor 0 si la furgoneta tiene motor de seis cilindros y el valor 1 si la furgoneta tiene motor de ocho cilindros. Obtenga la ecuación de regresión estimada que sirva para predecir el rendimiento de combustible en ciudad cuando se conoce el desplazamiento del motor y el valor de las variables ficticias Tracción4 y OchoCil.
e. Pruebe la significancia global y la significancia de cada una de las variables en la ecuación obtenida en el inciso d). Use $\alpha = 0.05$.

57. En el mercado actual se ofrece una amplia variedad de vehículos utilitarios deportivos o SUV (acrónimo en inglés de Sport Utility Vehicle) y de pickups. Para muchos de los compradores es un factor importante el valor de reventa del vehículo. En la tabla siguiente se presenta el valor de reventa (%) después de dos años y se sugiere el precio de 10 SUV, de 10 pickups pequeñas y de 10 pickups grandes (*Kipkinger's New Cars & Truckers 2000 Buyer's Guide*).



Trucks

	Tipo de vehículo	Precio sugerido ($)	Valor de reventa (%)
Chevrolet Blazer LS	utilitario deportivo	19 495	55
Ford Explorer Sport	utilitario deportivo	20 495	57
GMC Yukon XL 1500	utilitario deportivo	26 789	67
Honda CR-V	utilitario deportivo	18 965	65
Isuzu VehiCross	utilitario deportivo	30 186	62
Jeep Cherokee Limited	utilitario deportivo	25 745	57
Mercury Mountaineer	utilitario deportivo	29 895	59
Nissan Pathfinder XE	utilitario deportivo	26 919	54
Toyota 4Runner	utilitario deportivo	22 418	55
Toyota RAV4	utilitario deportivo	17 148	55
Chevrolet S-10 Extended Cab	pickup pequeña	18 847	46
Dodge Dakota Club Cab Sport	pickup pequeña	16 870	53
Ford Ranger XLT Regular Cab	pickup pequeña	18 510	48
Ford Ranger XLT Supercab	pickup pequeña	20 225	55
GMC Sonoma Regular Cab	pickup pequeña	16 938	44
Isuzu Hombre Spacecab	pickup pequeña	18 820	41
Mazda B4000 SE Cab Plus	pickup pequeña	23 050	51
Nissan Frontier XE Regular Cab	pickup pequeña	12 110	51
Toyota Tacoma Xtracab	pickup pequeña	18 228	49
Toyota Tacoma Xtracab V6	pickup pequeña	19 318	50
Chevrolet K2500	pickup grande	24 417	60
Chevrolet Silverado 2500 Ext	pickup grande	24 140	64
Dodge Ram 1500	pickup grande	17 460	54
Dodge Ram Quad Cab 2500	pickup grande	32 770	63
Dodge Ram Regular Cab 2500	pickup grande	23 140	59
Ford F150 XL	pickup grande	22 875	58
Ford F-350 Super Duty Crew Cab XL	pickup grande	34 295	64
GMC New Sierra 1500 Ext Cab	pickup grande	27 089	68
Toyota Tundra Access Cab Limited	pickup grande	25 605	53
Toyota Tundra Regular Cab	pickup grande	15 835	58

a. Obtenga la ecuación de regresión estimada que sirva para predecir el valor de reventa conociendo el precio sugerido. Pruebe la significancia de la relación empleando como nivel de significancia 0.05.
b. ¿Proporciona la ecuación de regresión estimada obtenida en el inciso a) un buen ajuste a los datos? Explique.
c. Obtenga la ecuación de regresión estimada que sirva para predecir el valor de reventa conociendo el precio sugerido y el tipo de vehículo.
d. Emplee la prueba F para determinar la significancia de los resultados de la regresión. ¿A qué conclusión se llega empleando 0.05 como nivel de significancia?

Caso problema 1 Consumer Research, Inc.

Consumer Research, Inc., es una empresa que realiza estudios para otras empresas sobre las actitudes y el comportamiento de los consumidores. En un estudio, el cliente solicitó un estudio sobre las características de los consumidores que pueden servir para predecir los montos que cargan a sus tarjetas de crédito. De una muestra de 50 consumidores se obtuvieron datos sobre ingreso anual, tamaño de la familia, y cargos anuales hechos las tarjetas de crédito. Los datos que se presentan a continuación se encuentran en el archivo Consumer del disco compacto que se distribuye con el libro.



Consumer

Ingreso (miles de $)	Tamaño de la familia	Monto de los cargos ($)	Ingreso (miles de $)	Tamaño de la familia	Monto de los cargos ($)
54	3	4016	54	6	5573
30	2	3159	30	1	2583
32	4	5100	48	2	3866
50	5	4742	34	5	3586
31	2	1864	67	4	5037
55	2	4070	50	2	3605
37	1	2731	67	5	5345
40	2	3348	55	6	5370
66	4	4764	52	2	3890
51	3	4110	62	3	4705
25	3	4208	64	2	4157
48	4	4219	22	3	3579
27	1	2477	29	4	3890
33	2	2514	39	2	2972
65	3	4214	35	1	3121
63	4	4965	39	4	4183
42	6	4412	54	3	3730
21	2	2448	23	6	4127
44	1	2995	27	2	2921
37	5	4171	26	7	4603
62	6	5678	61	2	4273
21	3	3623	30	2	3067
55	7	5301	22	4	3074
42	2	3020	46	5	4820
41	7	4828	66	4	5149

Reporte administrativo

1. Usando los métodos de la estadística descriptiva dé un resumen de estos datos. Haga un comentario sobre sus hallazgos.
2. Obtenga ecuaciones de regresión estimada, usando como variable independiente, primero, el ingreso anual y después, el tamaño de la familia. ¿Cuál de estas variables es mejor predictor de los cargos a las tarjetas de crédito? Analice sus hallazgos.
3. Obtenga una ecuación de regresión estimada en la que ingreso anual y tamaño de la familia sean las variables independientes. Analice sus hallazgos.
4. ¿Cuál es el monto del cargo anual en tarjetas de crédito que se puede predecir para un hogar de tres personas con ingreso anual de $4000?
5. Analice la necesidad de agregar otras variables independientes al modelo. ¿Qué variables sería útil agregar al modelo?

Caso problema 2 Predicción de la puntuación en un examen

Con objeto de predecir cómo se clasificarán los distritos escolares al considerar la pobreza y otras medidas relacionadas con el ingreso, *The Cincinnati Enquirer* reunió datos del Servicio de Administración de la Educación del Departamento de Educación de Ohio y del Departamento de Impuestos de Ohio (*The Cincinnati Enquirer*, 30 de noviembre de 1997). Primero, a principios de 1996, este periódico obtuvo datos sobre las calificaciones de aprobación en matemáticas, lectura, ciencias, escritura y civismo para alumnos de cuarto, noveno y doceavo grados. Combinando estos datos, el periódico calculó el porcentaje total de estudiantes que aprobaba los exámenes en cada distrito.

Se registró también el porcentaje de estudiantes del distrito escolar que recibía Apoyo para menores dependientes (o Aid for Dependent Children, ADC), el porcentaje de quienes calificaron para recibir almuerzos gratis o de precio reducido, así como el ingreso familiar mediano. Una fracción de los datos registrados para los 608 distritos escolares es la siguiente. (El conjunto completo de datos está disponible en el CD que acompaña al libro en el archivo llamado Enquirer).



Enquirer

Posición	Distrito escolar	Condado	% de aprobados	% con ADC	% con almuerzo gratis	Ingreso mediano ($)
1	Ottawa Hills Local	Lucas	93.85	0.11	0.00	48 231
2	Wyoming City	Hamilton	93.08	2.95	4.59	42 672
3	Oakwood City	Montgomery	92.92	0.20	0.38	42 403
4	Madeira City	Hamilton	92.37	1.50	4.83	32 889
5	Indian Hill Ex Vill	Hamilton	91.77	1.23	2.70	44 135
6	Solon City	Cuyahoga	90.77	0.68	2.24	34 993
7	Chagrin Falls Ex Vill	Cuyahoga	89.89	0.47	0.44	38 921
8	Mariemont City	Hamilton	89.80	3.00	2.97	31 823
9	Upper Arlington City	Franklin	89.77	0.24	0.92	38 358
10	Granville Ex Vill	Licking	89.22	1.14	0.00	36 235
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Los datos se han ordenado de acuerdo a los porcentajes en la columna % de aprobados, que son los porcentajes totales de estudiantes que aprueban los exámenes. Los datos en la columna titulada % con ADC son los porcentajes de estudiantes, de cada distrito escolar, que reciben ADC, y los datos en la columna cuyo encabezado es % con almuerzo gratis son los porcentajes de estudiantes acreedores de almuerzo gratuito o a precio reducido. En la columna cuyo encabezado es ingreso mediano, se da el ingreso mediano por familia en cada distrito. Para cada uno de los distritos escolares se da también el condado en el que ubica el distrito escolar. Obsérvese que hay casos en los que en la columna % de almuerzo gratuito aparece un 0, esto indica que ese distrito no participa en el programa de almuerzo gratuito.

Reporte administrativo

Analice este conjunto de datos empleando los métodos presentados en este capítulo y en los capítulos anteriores. Presente un resumen de su análisis comprendiendo los resultados estadísticos clave, las conclusiones, y las recomendaciones en un reporte administrativo. Incluya todo material técnico que considere sea adecuado en un apéndice.

Caso problema 3 Aportaciones de los alumnos

Las donaciones de los exalumnos son una importante fuente de ingresos para las universidades. Si los administradores pudieran determinar los factores que influyen sobre el aumento de la cantidad de alumnos que hacen donaciones, podrían poner en marcha políticas que llevarían a ganancias mayores. Las investigaciones indican que los estudiantes más satisfechos con la relación con sus profesores tienen más probabilidad de titularse, lo que a su vez lleva al aumento de la cantidad de alumnos que hagan donaciones. En la tabla 15.13 se muestran datos de 48 universidades de Estados Unidos (*American's Best Collage,* edición del año 2000). La columna titulada "tasa de titulados" da el porcentaje de alumnos titulados, de los inicialmente inscritos. La columna que tiene como título "% de grupos con menos de 20" da muestra el porcentaje de grupos con menos de 20 alumnos. La columna que tiene como título "Tasa de estudiantes/facultad" da la cantidad total de estudiantes inscritos, dividida entre la cantidad de facultades. Por último, la co-



Alumni

TABLA 15.13 DATOS DE 48 UNIVERSIDADES NACIONALES (DE ESTADOS UNIDOS)

	Estado	Tasa de titulados	% de grupos con menos de 20	Tasa de estudiantes/ facultad	Tasa de alumnos que donan
Boston College	MA	85	39	13	25
Brandeis University	MA	79	68	8	33
Brown University	RI	93	60	8	40
California Institute of Technology	CA	85	65	3	46
Carnegie Mellon University	PA	75	67	10	28
Case Western Reserve Univ.	OH	72	52	8	31
College of William and Mary	VA	89	45	12	27
Columbia University	NY	90	69	7	31
Cornell University	NY	91	72	13	35
Dartmouth College	NH	94	61	10	53
Duke University	NC	92	68	8	45
Emory University	GA	84	65	7	37
Georgetown University	DC	91	54	10	29
Harvard University	MA	97	73	8	46
Johns Hopkins University	MD	89	64	9	27
Lehigh University	PA	81	55	11	40
Massachusetts Inst. of Technology	MA	92	65	6	44
New York University	NY	72	63	13	13
Northwestern University	IL	90	66	8	30
Pennsylvania State Univ.	PA	80	32	19	21
Princeton University	NJ	95	68	5	67
Rice University	TX	92	62	8	40
Stanford University	CA	92	69	7	34
Tufts University	MA	87	67	9	29
Tulane University	LA	72	56	12	17
U. of California–Berkeley	CA	83	58	17	18
U. of California–Davis	CA	74	32	19	7
U. of California–Irvine	CA	74	42	20	9
U. of California–Los Angeles	CA	78	41	18	13
U. of California–San Diego	CA	80	48	19	8
U. of California–Santa Barbara	CA	70	45	20	12
U. of Chicago	IL	84	65	4	36
U. of Florida	FL	67	31	23	19
U. of Illinois–Urbana Champaign	IL	77	29	15	23
U. of Michigan–Ann Arbor	MI	83	51	15	13
U. of North Carolina–Chapel Hill	NC	82	40	16	26
U. of Notre Dame	IN	94	53	13	49
U. of Pennsylvania	PA	90	65	7	41
U. of Rochester	NY	76	63	10	23
U. of Southern California	CA	70	53	13	22
U. of Texas–Austin	TX	66	39	21	13
U. of Virginia	VA	92	44	13	28
U. of Washington	WA	70	37	12	12
U. of Wisconsin–Madison	WI	73	37	13	13
Vanderbilt University	TN	82	68	9	31
Wake Forest University	NC	82	59	11	38
Washington University–St. Louis	MO	86	73	7	33
Yale University	CT	94	77	7	50

lumna que tiene como título "Tasa de alumnos que donan" da el porcentaje de alumnos que han hecho alguna donación a la universidad.

Reporte administrativo

1. Resuma estos datos empleando los métodos de la estadística descriptiva.
2. Obtenga una ecuación de regresión estimada que sirva para predecir la tasa de donaciones de los alumnos dada la cantidad de alumnos que se titulan. Analice sus hallazgos.
3. Obtenga una ecuación de regresión estimada que sirva para predecir la tasa de donaciones de los alumnos empleando los datos proporcionados.
4. ¿Qué conclusiones y recomendaciones puede obtener de su análisis?

Caso problema 4 Predicción del porcentaje de triunfos de la NFL



NFLStats

En la liga nacional de futbol americano de Estados Unidos se da seguimiento a diversos datos de desempeño tanto del equipo como de los jugadores (www.nfl.com). En el archivo titulado NLFStats del disco compacto se presentan algunos de los datos de fin de año de la temporada del 2005 de la NFL. Cada renglón corresponde a los datos de un equipo de la NFL, y los equipos aparecen de acuerdo a su porcentaje de juegos ganados. A continuación se da la descripción de los datos.

WinPct	Porcentaje de juegos ganados
DefYds/G	Promedio de yardas permitidas por la defensiva del juego
RushYds/G	Promedio de yardas por carrera
PassYds/G	Promedio de yardas por aire por juego
FGPct	Porcentaje de goles de campo
TakeInt	TakeInt Intercepciones obtenidas; total de intercepciones obtenidas por la defensa del equipo
TakeFum	TakeFun Balones sueltos obtenidos; total de balones recuperados por la defensa del equipo
GiveInt	GiveInt Intercepciones otorgadas; total de intercepciones hechas por la ofensiva del equipo
GiveFum	GiveFun Balones sueltos perdidos; total de balones perdidos por la defensa del equipo

Reporte administrativo

1. Resuma los datos empleando los métodos de la estadística descriptiva. Haga un comentario sobre sus hallazgos.
2. Obtenga una ecuación de regresión estimada que sirva para predecir WinPCT usando DefYds/G, RushYds/G, PassYds/G y FGPct. Analice sus hallazgos.
3. En la ecuación de regresión estimada obtenida en el inciso 2), elimine todas las variables independientes que no sean significativas y obtenga otra ecuación de regresión estimada que sirva para predecir WinPct. Use $\alpha = 0.05$.
4. Algunos analistas de futbol americano consideran que las pérdidas de balón son uno de los factores más importantes para determinar el éxito de un equipo. Si Takeaways = TakeInt + TakeFum y Giveayaws = GiveInt + GiveFum, sea NetDiff = Takeaways − Giveaways. Obtenga una ecuación de regresión estimada que sirva para predecir WinPct empleando NetDiff. Compare estos resultados con la ecuación de regresión estimada obtenida en el inciso 3).
5. Obtenga una ecuación de regresión estimada que permita predecir WinPct usando todos los datos proporcionados.

Apéndice 15.1 Regresión múltiple con Minitab



Butler

En la sección 15.2 se vio la solución a problemas de regresión múltiple mediante paquetes de software mediante los resultados dados por Minitab al problema de la empresa Butler Trucking. En este apéndice se describen los pasos requeridos para que Minitab genere esos resultados. Primero, es necesario ingresar los datos en una hoja de cálculo de Minitab. Las millas recorridas se ingresan en la columna C1, el número de entregas se ingresan en la columna C2 y el tiempo de recorrido (en horas) en la columna C3. Los nombres de las variables, Miles (millas), Deliveries (entregas) y Time (tiempo) se ingresan como encabezados de estas columnas. En la explicación siguiente se hará referencia a los datos empleando los nombres de las variables o los identificadores de las columnas C1, C2 y C3. A continuación se describen los pasos a seguir con Minitab para obtener los resultados de regresión que se presentan en la figura 15.4.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Seleccionar el menú **Regression**
Paso 3. Elegir **Regression**
Paso 4. Cuando aparezca el cuadro de diálogo **Regression**
Ingresar Time en la caja **Response**
Ingresar Miles y Deliveries en el cuadro **Predictors**
Clic en **OK**

Apéndice 15.2 Regresión múltiple con Excel

archivo en CD

Butler

En la sección 15.2 se vio la solución a problemas de regresión múltiple empleando paquetes de software mediante los resultados dados por Minitab al problema de la empresa Butler Trucking. En este apéndice se describe el uso de la herramienta de regresión de Excel para obtener la ecuación de regresión estimada para el problema de Butler Trucking. A medida que se describen los pasos a seguir consúltese la figura 15.14. Primero, en las celdas A1:D1 de la hoja de cálculo se ingresan los rótulos Recorrido, Millas, Entregas y Tiempo, y en las celdas B2:D11 se ingresan los datos muestrales. En las celdas A2:A11, los números 1-10 sirven para identificar cada observación.




Los pasos siguientes describen cómo emplear la herramienta de regresión para el análisis de regresión múltiple.

Paso 1. Seleccionar el menú **Herramientas**
Paso 2. Elegir **Análisis de datos**
Paso 3. Elegir **Regresión** en la lista Funciones para análisis
Paso 4. Cuando aparezca el cuadro de diálogo Regresión
Ingresar D1:D11 en el cuadro **Rango Y de entrada**
Ingresar B1:C11 en el cuadro **Rango X de entrada**
Seleccionar **Rótulos**
Seleccionar **Nivel de confianza**
Ingresar 99 en el cuadro **Nivel de confianza**
Seleccionar **Rango de salida**
Ingresar A13 en el cuadro **Rango de salida** (para identificar la esquina superior izquierda de la de la hoja de cálculo en donde aparecerán los resultados)
Clic en **Aceptar**

En los resultados de Excel que se presentan en la figura 15.14, el rótulo para la variable independiente x_1 es Millas (ver celda A30) y el rótulo para la variable independiente x_2 es Entregas (ver celda A31). La ecuación de regresión estimada es

$$\hat{y} = -0.8687 + 0.0611x_1 + 0.9234x_2$$

FIGURA 15.14 RESULTADOS DADOS POR EXCEL AL PROBLEMA DE BUTLER TRUCKING CON DOS VARIABLES INDEPENDIENTES

	A	B	C	D	E	F	G	H	I	J
1	**Assignment**	**Miles**	**Deliveries**	**Time**						
2	1	100	4	9.3						
3	2	50	3	4.8						
4	3	100	4	8.9						
5	4	100	2	6.5						
6	5	50	2	4.2						
7	6	80	2	6.2						
8	7	75	3	7.4						
9	8	65	4	6						
10	9	90	3	7.6						
11	10	90	2	6.1						
12										
13	SUMMARY OUTPUT									
14										
15	*Regression Statistics*									
16	Multiple R	0.9507								
17	R Square	0.9038								
18	Adjusted R Square	0.8763								
19	Standard Error	0.5731								
20	Observations	10								
21										
22	ANOVA									
23		*df*	*SS*	*MS*	*F*	*Significance F*				
24	Regression	2	21.6006	10.8003	32.8784	0.0003				
25	Residual	7	2.2994	0.3285						
26	Total	9	23.9							
27										
28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*	
29	Intercept	-0.8687	0.9515	-0.9129	0.3916	-3.1188	1.3813	-4.1986	2.4612	
30	Miles	0.0611	0.0099	6.1824	0.0005	0.0378	0.0845	0.0265	0.0957	
31	Deliveries	0.9234	0.2211	4.1763	0.0042	0.4006	1.4463	0.1496	1.6972	
32										

Obsérvese que el uso de la herramienta de regresión de Excel para la regresión múltiple es casi igual a su uso para regresión lineal simple. La principal diferencia es que en el caso de la regresión múltiple se requiere un rango mayor de celdas para identificar las variables independientes.

Apéndice 15.3 Regresión logística con Minitab



Simmons

Minitab le llama a la regresión logística con una sola variable independiente que puede tomar los valores 0 y 1, Regresión Logística Binaria (Binary Logistic Regresion). En este apéndice se describen los pasos que se requieren en el procedimiento de regresión logística binaria de Minitab para generar los resultados presentados en la figura 15.13 para el problema de Simmons. Primero, en una hoja de cálculo de Minitab deben ingresarse los datos. Las cantidades (en miles de $) que gastaron los clientes en las tiendas Simmons se ingresan en la columna C2, los datos sobre la tarjeta de crédito (1 si tienen tarjeta de crédito de Simmons; 0 si no es así) se ingresan en la columna C3 y el dato sobre el uso del cupón (1 si el cliente usó el cupón; 0 si no fue así) se ingresan en la columna C4. Los nombres de las variables Spending (gasto) Card (tarjeta) y Coupon (cupón) se ingresan en la hoja de cálculo como encabezados de las columnas.

En la explicación siguiente se hará referencia a los datos empleando los nombres de las variables o los identificadores de las columnas C2, C3 y C4. Mediante los pasos siguientes se generarán los resultados de la regresión logística.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Seleccionar el menú **Regression**
Paso 3. Elegir **Binary Logistic Regression**
Paso 4. Cuando aparezca el cuadro de diálogo **Binary Logistic Regression**
Ingresar Coupon en el cuadro **Response**
Ingresar Spending y Card en el cuadro **Model**
Clic en **OK**

La información presentada en la figura 15.13 aparecerá como parte de los resultados.

CAPÍTULO 16



Análisis de regresión: construcción de modelos

CONTENIDO



LA ESTADÍSTICA *en* LA PRÁCTICA

LA EMPRESA MONSANTO*
SAN LUIS MISSOURI

Las raíces de Monsanto se remontan a una inversión de $500 de un empresario y a un polvoriento almacén a orillas del Mississippi, donde en 1901 John F. Queeney empezó con la fabricación de sacarina. En la actualidad, Monsanto es una de las mayores empresas químicas de Estados Unidos, produce más de mil productos que van desde productos químicos industriales hasta canchas sintéticas para deportes que se emplean en los estadios modernos. Monsanto es una empresa mundial que cuenta con fábricas, laboratorios, centros técnicos y operaciones de marketing en 65 países.

La división de nutrición química de Monsanto fabrica y comercializa un suplemento de metionina que se usa en alimento para ganado, cerdos y aves de corral. Como en la cría de aves de corral se trabaja con volúmenes altos y márgenes de ganancias reducidos, se necesitan alimentos rentables y con el mayor valor nutricional posible. El alimento de composición óptima es el que produce un crecimiento rápido y un alto peso corporal final con una determinada ingestión de alimento. La industria química trabaja en estrecha colaboración con los criadores de aves de corral para optimizar los alimentos. Por último, el éxito depende de mantener bajo el costo de las aves de corral en comparación con el costo de la carne de res y de otros productos de carne.

Para modelar la relación entre peso corporal y y cantidad de metionina x adicionada al alimento para aves de corral, los investigadores de Monsanto emplearon el análisis de regresión. Al principio se obtuvo la siguiente ecuación de regresión lineal estimada.

$$\hat{y} = 0.21 + 0.42x$$

Esta ecuación estimada de regresión resultó estadísticamente significativa; sin embargo, de acuerdo con el análisis de residuales una relación curvilínea parecía ser un modelo más adecuado para la relación entre peso corporal y metionina.

*Los autores agradecen a James R. Ryland y a Robert M. Schisla, especialistas en investigación de Monsanto Nutrition Chemical Division, por proporcionar este artículo para *La estadística en la práctica*.



Los investigadores de Monsanto emplearon el análisis de regresión para obtener un alimento de composición óptima para los criadores de aves de corral. © PhotoDisc/Getty Images.

Al continuar con las investigaciones encontraron que aunque cantidades pequeñas de metionina tendían a hacer aumentar el peso corporal, en cierto punto el peso corporal se estabilizaba y un aumento en la cantidad de metionina tenía poco o ningún efecto. Peor aún, cuando la cantidad de metionina era mayor que el requerimiento nutrimental, el peso corporal tendía a disminuir. Para modelar la relación curvilínea entre peso corporal y metionina se empleó la siguiente ecuación estimada de regresión múltiple.

$$\hat{y} = -1.89 + 1.32x - 0.506x^2$$

Al aplicar la regresión, los investigadores de Monsanto pudieron encontrar la cantidad óptima de metionina que debía usarse en los productos alimenticios para aves de corral.

En este capítulo el estudio del análisis de regresión se ampliará a la obtención de modelos curvilíneos como el usado por los investigadores de Monsanto. Se describirán, además, diversas herramientas que sirven para determinar cuáles son las variables independientes que conducen a una mejor ecuación estimada de regresión.

La construcción de modelos es el proceso que consiste en obtener una ecuación estimada de regresión que describa la relación entre una variable dependiente y una o diversas variables independientes. Lo más importante en la construcción de un modelo es hallar la forma funcional adecuada para la relación y seleccionar las variables independientes que se deban incluir en el modelo. En la sección 16.1 se presenta el concepto de modelo lineal general que establece el marco para la construcción de modelos. En la sección 16.2, en la que se presentan las bases para procedimientos más sofisticados que emplean paquetes de software, se enseña un procedimiento

general para determinar cuándo agregar o eliminar variables independientes. En la sección 16.3 se considera un problema más grande de regresión en el que intervienen ocho variables independientes y 25 observaciones; este problema sirve para ilustrar los procedimientos de selección de variables presentados en la sección 16.4, que comprenden la regresión por pasos, el procedimiento de selección hacia adelante, el procedimiento de eliminación hacia atrás y el mejor subconjunto de regresión. En la sección 16.5 se muestra cómo el análisis de regresión múltiple proporciona otro método para la solución de problemas de diseño experimental, y en la sección 16.6 se muestra el uso de la prueba Durbin-Watson para detectar correlación serial o autocorrelación.

16.1 El modelo lineal general

Suponga que se obtienen los datos de una variable independiente y y de k variables independientes $x_1, x_2, \ldots, x_k$. El objetivo es obtener, con estos datos, la ecuación estimada de regresión que mejor exprese la relación entre la variable dependiente y las independientes. Como marco general para el desarrollo de relaciones más complejas entre las variables independientes, se introduce el concepto de **modelo lineal general** con p variables independientes.

Si el modelo de regresión se puede expresar en la forma de la ecuación (16.1), entonces se aplica el procedimiento estándar de regresión múltiple descrito en el capítulo 15.

MODELO LINEAL GENERAL

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \epsilon \qquad \textbf{(16.1)}$$

En la ecuación (16.1), cada una de las variables independientes z_j (donde $j = 1, 2, \ldots, p$) es función de $x_1, x_2, \ldots, x_k$ (las variables para las que se obtuvieron los datos). En algunos casos cada z_j puede ser función de una sola variable x. El caso más sencillo es cuando sólo se obtienen datos de una variable x_1 y se quiere estimar y por medio de una relación lineal. En este caso $z_1 = x_1$ y la ecuación (16.1) se convierte en

$$y = \beta_0 + \beta_1 x_1 + \epsilon \qquad \textbf{(16.2)}$$

La ecuación (16.2) es el modelo de regresión lineal simple presentado en el capítulo 14, con la única diferencia de que a la variable independiente se le ha llamado x_1 en lugar de x. En la literatura sobre modelos estadísticos, a este modelo se le llama *modelo simple de primer orden con una variable predictora.*

TABLA 16.1

DATOS DEL EJEMPLO DE REYNOLDS

Antigüedad en meses	Balanzas vendidas
41	375
106	296
76	317
10	376
22	162
12	150
85	367
111	308
40	189
51	235
9	83
12	112
6	67
56	325
19	189



Modelado de relaciones curvilíneas

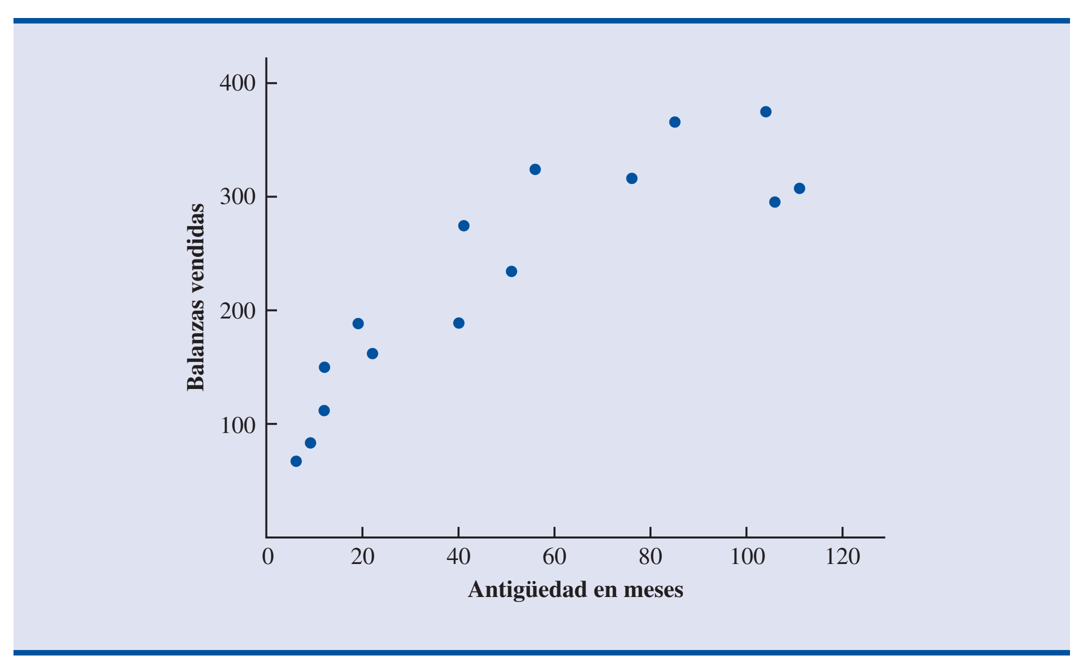
Con la ecuación (16.1) también se pueden modelar relaciones más complejas. Para ilustrar esto se verá un problema que se le presentó a la empresa Reynolds, Inc., fabricante de balanzas industriales y de equipo para laboratorio. Los gerentes de Reynolds desean investigar la relación que existe entre la antigüedad de sus vendedores y el número de balanzas electrónicas para laboratorio que venden. En la tabla 16.1 se presenta el número de balanzas vendidas por cada uno de 15 vendedores elegidos aleatoriamente y la antigüedad que tiene cada uno de ellos en la empresa. En la figura 16.1 se presenta el diagrama de dispersión de estos datos. En el diagrama de dispersión se observa que es posible que exista una relación curvilínea entre antigüedad de un empleado y número de balanzas que vende. Antes de considerar cómo obtener una relación curvilínea para este problema de Reynolds, se analizarán los resultados de Minitab que se presentan en la figura 16.2 y que corresponden a un modelo simple de primer orden; la ecuación estimada de regresión es

$$\text{Sales (Ventas)} = 111 + 2.38 \text{ Months (Meses)}$$

donde

Sales (Ventas) = número de balanzas electrónicas para laboratorio vendidas
Months (Meses) = antigüedad del vendedor, en meses

FIGURA 16.1 DIAGRAMA DE DISPERSIÓN PARA EL EJEMPLO DE REYNOLDS



La figura 16.3 es la gráfica de residuales estandarizados correspondiente. Aunque los resultados de Minitab indican que la relación sí es significativa (valor-p = 0.000) y que una relación lineal explica un porcentaje grande de la variabilidad en las ventas (R-sq = 78.1%), la gráfica de residuales estandarizados sugiere que se necesita una relación curvilínea.

Para obtener una relación curvilínea, en la ecuación (16.1) se hace $z_1 = x_1$ y $z_2 = x_1^2$, así resulta el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon \tag{16.3}$$

A este modelo se le llama *modelo de segundo orden con una variable predictora.* Para proporcionar la ecuación estimada de regresión correspondiente a este modelo de segundo orden, Minitab

FIGURA 16.2 RESULTADOS DE MINITAB PARA EL EJEMPLO DE REYNOLDS: MODELO DE PRIMER ORDEN.

```
The regression equation is
Sales = 111 + 2.38 Months

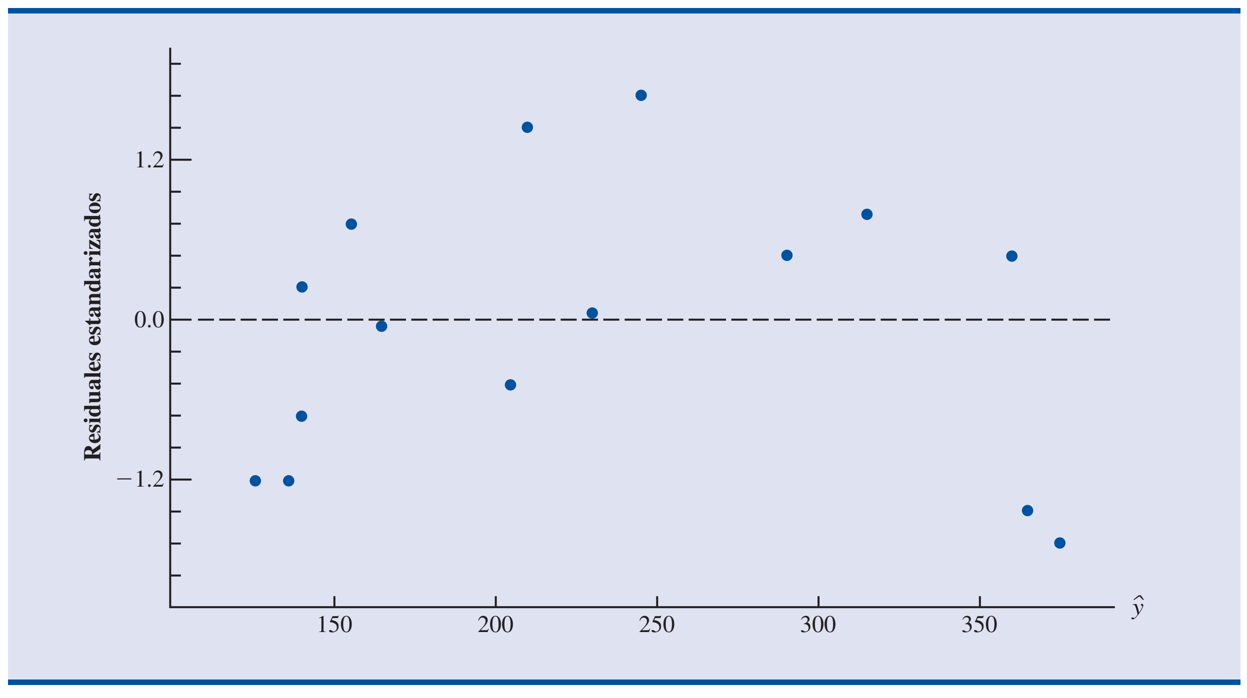
Predictor    Coef   SE Coef     T      p
Constant   111.23     21.63  5.14  0.000
Months     2.3768    0.3489  6.81  0.000

S = 49.52   R-sq = 78.1%   R-sq(adj) = 76.4%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       1  113783  113783  46.41  0.000
Residual Error  13   31874    2452
Total           14  145657
```

FIGURA 16.3 GRÁFICA DE RESIDUALES ESTANDARIZADOS DEL EJEMPLO DE REYNOLDS: MODELO DE PRIMER ORDEN



necesita los datos originales de la tabla 16.1, así como los datos que corresponden a la segunda variable dependiente que se agrega, que es el cuadrado de los meses de antigüedad que tiene el empleado en la empresa. En la figura 16.4 se presentan los resultados de Minitab correspondientes al modelo de segundo orden; la ecuación estimada de regresión es

Los datos de la variable independiente MonthsSq *se obtienen al elevar al cuadrado los valores de* Months.

$$\text{Sales (Ventas)} = 45.3 + 6.34 \text{ Months (Meses)} - 0.0345 \text{ MonthsSq (Meses al cuadrado)}$$

donde

$$\text{MonthsSq} = \text{cuadrado del número de meses que ha trabajado el vendedor en la empresa}$$

La figura 16.5 es la gráfica de residuales estandarizados correspondiente. En esta gráfica se observa que el patrón curvilíneo anterior ha desaparecido. Al emplear como nivel de significancia 0.05, los resultados de Minitab indican que el modelo general es significativo (el valor-*p* para la prueba *F* es 0.000); observe también que el valor-*p* correspondiente al cociente-*t* de MonthsSq (valor-*p* = 0.002) es menor que 0.05, por lo que se puede concluir que la adición de MonthsSq al modelo es significativa. Como el valor de R-sq(adj) es 88.6%, se puede estar satisfecho con el ajuste que proporciona esta ecuación estimada de regresión. Lo más importante, sin embargo, es ver lo fácil que es tratar las relaciones curvilíneas en el análisis de regresión.

Es claro que por medio de la ecuación (16.1) se pueden modelar muchos tipos de relaciones. Las técnicas de regresión con las que se ha estado trabajando son técnicas que definitivamente no están limitadas a relaciones lineales. En el análisis de regresión múltiple, la palabra *lineal* en el término “modelo lineal general” se refiere únicamente al hecho de que $\beta_0, \beta_1, \ldots, \beta_p$, tienen, todos, exponente 1; no implica que la relación entre y y las x_i sea lineal. Es más, en esta sección se ha visto un ejemplo del uso de la ecuación (16.1) para modelar una relación curvilínea.

FIGURA 16.4 RESULTADOS DE MINITAB PARA EL EJEMPLO DE REYNOLDS: MODELO DE SEGUNDO ORDEN

```
The regression equation is
Sales = 45.3 + 6.34 Months - 0.0345 MonthsSq

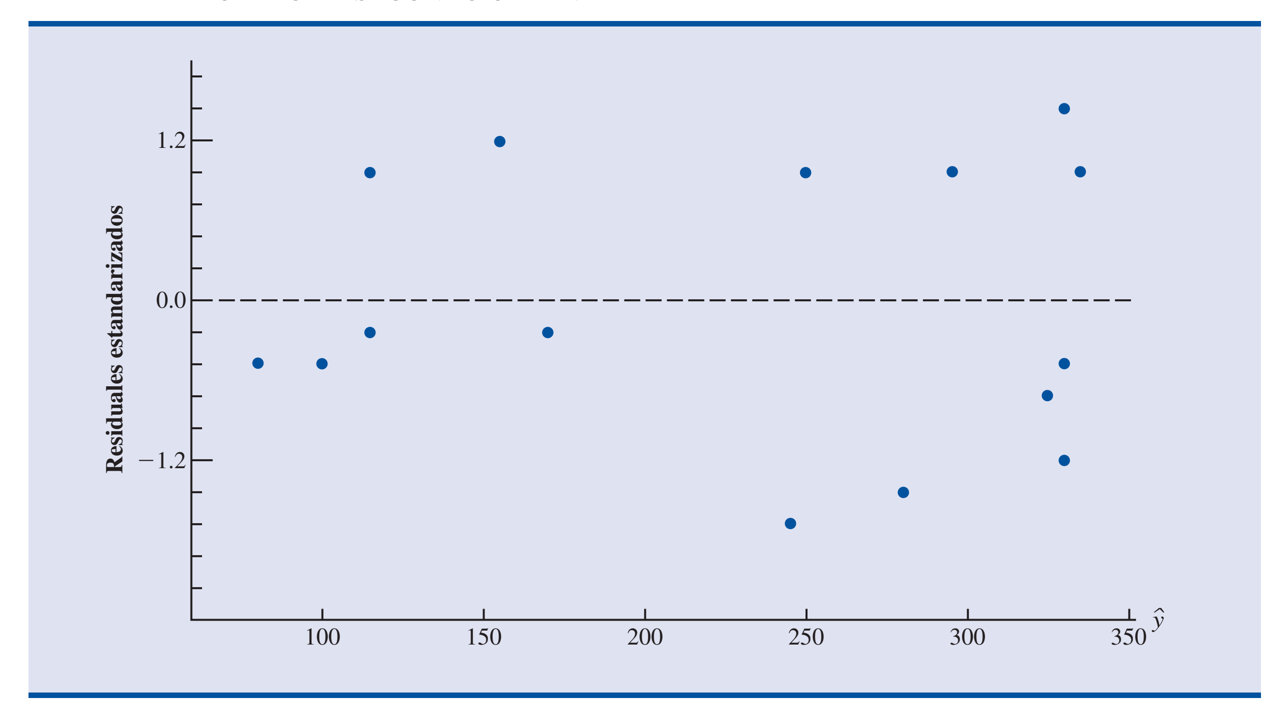
Predictor        Coef    SE Coef       T      p
Constant        45.35      22.77    1.99  0.070
Months          6.345      1.058    6.00  0.000
MonthsSq    -0.034486   0.008948   -3.85  0.002

S = 34.45   R-sq = 90.2%   R-sq(adj) = 88.6%

Analysis of Variance

SOURCE          DF      SS      MS       F      p
Regression       2  131413   65707   55.36  0.000
Residual Error  12   14244    1187
Total           14  145657
```

FIGURA 16.5 GRÁFICA DE RESIDUALES ESTANDARIZADOS PARA EL EJEMPLO DE REYNOLDS: MODELO DE SEGUNDO ORDEN



Interacción

Si el conjunto de datos original consta de observaciones para y y para dos variables independientes x_1 y x_2, y en la ecuación (16.1) se pone $z_1 = x_1$, $z_2 = x_2$, $z_3 = x_1^2$, $z_4 = x_2^2$ y $z_5 = x_1x_2$ se puede obtener un modelo de segundo orden con dos variables predictoras. El modelo resultante es

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1^2 + \beta_4x_2^2 + \beta_5x_1x_2 + \epsilon \tag{16.4}$$

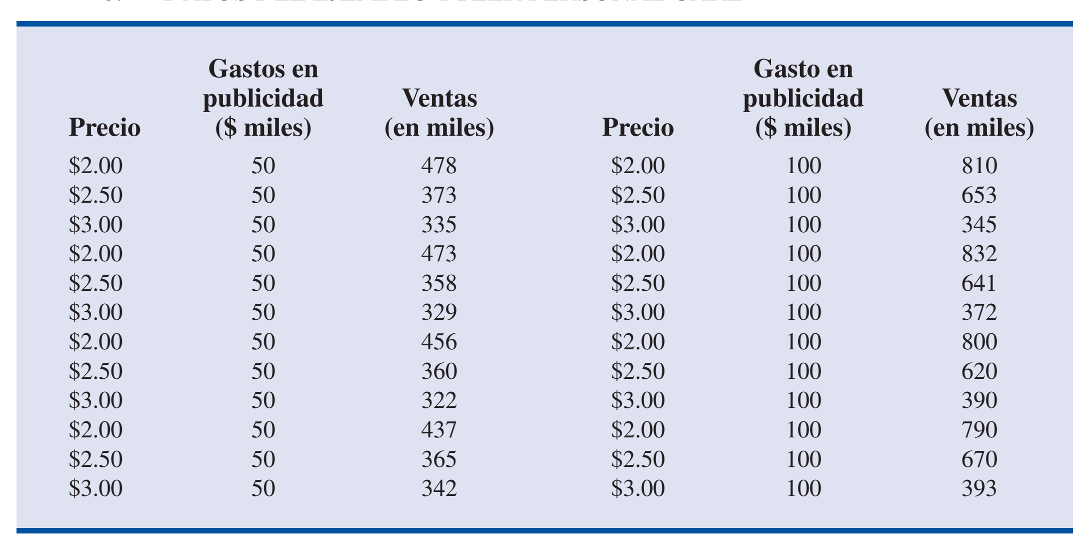
En este modelo de segundo orden, la variable $z_5 = x_1x_2$ se agrega para tomar en cuenta el posible efecto que pueda tener la acción conjunta de las dos variables. A este tipo de efecto se le llama **interacción**.

Para ver un ejemplo de interacción y de lo que significa, se revisará el estudio de regresión realizado por Tyler Personal Care para un nuevo champú. Se pensó que los dos factores que tenían más influencia sobre las ventas eran el precio de venta por unidad y gasto en publicidad. Para investigar el efecto de estas dos variables sobre las ventas, se formaron parejas con los precios \$2.00, \$2.50 y \$3.00 y los gastos en publicidad \$50 000 y \$100 000 en 24 mercados de prueba. En la tabla 16.2 se presentan las unidades vendidas (en miles).

En la tabla 16.3 se resumen estos datos. Observe que las ventas medias muestrales correspondientes al precio \$2.00 y al gasto en publicidad \$50 000 son 461 000 unidades y que las ventas medias muestrales correspondientes al precio \$2.00 y al gasto en publicidad \$100 000 son 808 000 unidades. Por tanto, cuando el precio se mantiene constante en \$2.00, la diferencia en ventas medias entre los gastos en publicidad de \$50 000 y de \$100 000 es 808 000 − 461 000 = 347 000 unidades. Cuando el precio del producto es \$2.50, la diferencia en las ventas medias es 646 000 − 364 000 = 282 000 unidades. Por último, cuando el precio es \$3.00 la diferencia en las ventas medias es 375 000 − 332 000 = 43 000 unidades. Es claro que la diferencia en ventas medias entre gastos en publicidad de \$50 000 y de \$100 000 depende del precio del producto. En otras palabras, a precios de venta más elevados, el efecto del aumento en los gastos en publicidad disminuye. Estas observaciones hacen evidente la interacción entre las variables precio y gasto en publicidad.

Para proporcionar otra perspectiva de la interacción, en la figura 16.6 se muestran las ventas medias muestrales correspondientes a las seis combinaciones precio-gasto en publicidad. En esta gráfica también se muestra que el efecto de los gastos en publicidad sobre las ventas medias de-

TABLA 16.2 DATOS DEL EJEMPLO TYLER PERSONAL CARE

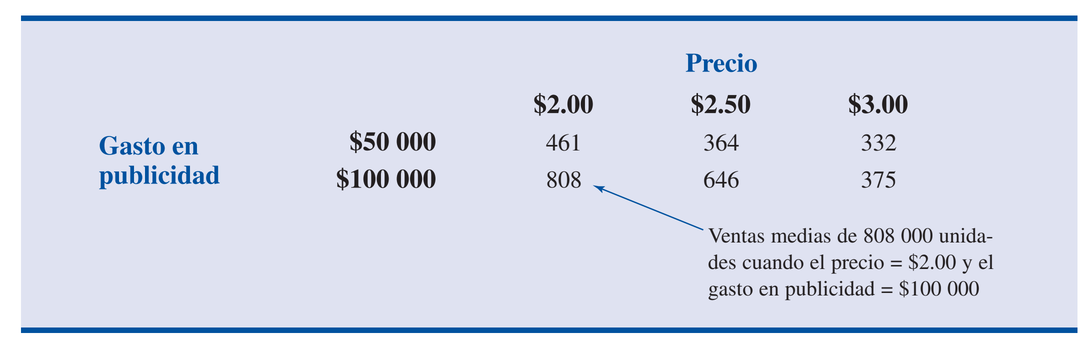


Precio	Gastos en publicidad ($ miles)	Ventas (en miles)	Precio	Gasto en publicidad ($ miles)	Ventas (en miles)
$2.00	50	478	$2.00	100	810
$2.50	50	373	$2.50	100	653
$3.00	50	335	$3.00	100	345
$2.00	50	473	$2.00	100	832
$2.50	50	358	$2.50	100	641
$3.00	50	329	$3.00	100	372
$2.00	50	456	$2.00	100	800
$2.50	50	360	$2.50	100	620
$3.00	50	322	$3.00	100	390
$2.00	50	437	$2.00	100	790
$2.50	50	365	$2.50	100	670
$3.00	50	342	$3.00	100	393



Tyler

TABLA 16.3 VENTAS UNITARIAS MEDIAS EN EL EJEMPLO DE TYLER PERSONAL CARE



		Precio		
		$2.00	**$2.50**	**$3.00**
Gasto en publicidad	**$50 000**	461	364	332
	$100 000	808	646	375

Ventas medias de 808 000 unidades cuando el precio = $2.00 y el gasto en publicidad = $100 000

FIGURA 16.6 VENTAS UNITARIAS MEDIAS (EN MILES) COMO FUNCIÓN DEL PRECIO DE VENTA Y DEL GASTO EN PUBLICIDAD

Ventas unitarias medias (en miles)
900
800
700
600
500
400
300
$100 000
$50 000
La diferencia es 808 – 461 = 347
$100 000
$50 000
La diferencia es 646 – 364 = 282
$100 000
$50 000
La diferencia es 375 – 332 = 43
2.00
2.50
3.00
Precio de venta ($)

pende del precio del producto; una vez más se ve el efecto de la interacción. Cuando hay interacción entre dos variables, no se puede estudiar el efecto de una de las variables sobre la respuesta y independientemente de la otra variable. En otras palabras, sólo es posible obtener conclusiones claras si se considera el efecto conjunto que tienen las dos variables sobre la respuesta.

Para tomar en cuenta el efecto de la interacción, se usará el siguiente modelo de regresión

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \quad \textbf{(16.5)}$$

donde

$$y = \text{ventas unitarias (en miles)}$$
$$x_1 = \text{precio (\$)}$$
$$x_2 = \text{gastos en publicidad (\$ miles)}$$

Observe que la ecuación (16.5) refleja la creencia de Tyler de que el número de unidades vendidas depende linealmente del precio de venta y de los gastos en publicidad (representados por los términos $\beta_1 x_1$ y $\beta_2 x_2$) y de que existe interacción entre las dos variables (representada por el término $\beta_3 x_1 x_2$).

Para obtener una ecuación estimada de regresión, se empleó un modelo lineal general con tres variables independientes (z_1, z_2 y z_3).

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon \quad \textbf{(16.6)}$$

donde

$$z_1 = x_1$$
$$z_2 = x_2$$
$$z_3 = x_1 x_2$$

En la figura 16.7 se presentan los resultados de Minitab correspondientes al modelo de interacción para el ejemplo de Tyler Personal Care. La ecuación estimada de regresión que se obtiene es

Sales (Ventas) = −276 + 175 Price (Precio) +
19.7 AdvExp (GastPubl) − 6.08 PriceAdv (PrecioPubl)

donde

Los datos de la variable independiente PriceAdv *se obtienen multiplicando cada precio por el correspondiente valor de* AdvExp.

Sales (Ventas) = ventas unitarias (miles)
Price (Precio) = precio del producto ($)
AdvExp (GastPubl) = gastos en publicidad ($ miles)
PriceAdv (PrecioPubl) = término de la interacción (precio multiplicado por publicidad)

Como el modelo es significativo (el valor-p en la prueba F es 0.000) y como el valor-p correspondiente a la prueba t para PriceAdv es 0.000, se concluye que la interacción es significativa dado el efecto lineal del precio del producto y de los gastos en publicidad. Por tanto, los resultados de la regresión indican que el efecto de gastos en publicidad sobre las ventas depende del precio.

Transformaciones a la variable dependiente

Al mostrar el uso del modelo lineal general para modelar diversas relaciones que puede haber entre las variables independientes y la variable dependiente, se ha concentrado la atención en transformaciones a una o varias de las variables independientes. Con frecuencia vale la pena utilizar

FIGURA 16.7 RESULTADOS DE MINITAB PARA EL EJEMPLO DE TYLER PERSONAL CARE

```
The regression equation is
Sales = - 276 + 175 Price + 19.7 AdvExpen - 6.08 PriceAdv

Predictor      Coef  SE Coef       T      p
Constant     -275.8    112.8   -2.44  0.024
Price        175.00    44.55    3.93  0.001
Adver        19.680    1.427   13.79  0.000
PriceAdv    -6.0800   0.5635  -10.79  0.000

S = 28.17   R-sq = 97.8%   R-sq(adj) = 97.5%

Analysis of Variance

SOURCE          DF      SS      MS       F      p
Regression       3  709316  236439  297.87  0.000
Residual Error  20   15875     794
Total           23  725191
```

TABLA 16.4

RENDIMIENTO EN MILLAS POR GALÓN Y PESOS DE 12 AUTOMÓVILES

Pesos	Millas por galón
2289	28.7
2113	29.2
2180	34.2
2448	27.9
2026	33.3
2702	26.4
2657	23.9
2106	30.5
3226	18.1
3213	19.5
3607	14.3
2888	20.9



transformaciones a la variable dependiente y. Para dar un ejemplo de un caso en el que puede ser útil transformar la variable dependiente, considere los datos de la tabla 16.4, en las que se muestran rendimientos en millas por galón y pesos de 12 automóviles. El diagrama de dispersión de la figura 16.8 indica que entre estas dos variables existe una relación lineal negativa. Por tanto, se usa un modelo simple de primer orden para relacionar estas dos variables. En la figura 16.9 se presentan los resultados que proporciona Minitab; la ecuación estimada de regresión es

$$\text{MPG} = 56.1 - 0.0116 \text{ Weight}$$

donde

MPG = rendimiento en millas por galón

Weight (Peso) = peso del automóvil dado en libras

El modelo es significativo (el valor-p en la prueba F es 0.000) y el ajuste es muy bueno (R-sq = 93.5%). Sin embargo, en la figura 16.9 se ve que la observación 3 ha sido identificada como una observación cuyo residual estandarizado es grande.

La figura 16.10 es la gráfica de los residuales estandarizados correspondientes al modelo de primer orden. Su forma no parece ser la de la banda horizontal que se esperaría observar si las suposiciones acerca del término del error fueran válidas. La variabilidad de los residuales parece aumentar a medida que aumenta el valor de $\hat{y}$. En otras palabras, se observa la forma de cuña que, como se dijo en los capítulos 14 y 15, indica una varianza que no es constante. Si las suposiciones para el modelo de esta prueba no parecen satisfacerse, entonces no se justifica sacar conclusiones acerca de la significancia estadística de la ecuación estimada de regresión que se obtiene.

El problema de una varianza no constante suele corregirse al transformar la variable dependiente a otra escala. Por ejemplo, si se trabaja con el logaritmo de la variable dependiente en lugar de la variable dependiente original, los valores de la variable dependiente se comprimirán (estarán más cercanos unos a otros) y con esto disminuirán los efectos de la varianza no constante. La mayor parte de los paquetes de software para estadística proporcionan la posibilidad de aplicar transformaciones logarítmicas mediante logaritmos base 10 (logaritmos comunes) o lo-

FIGURA 16.8 DIAGRAMA DE DISPERSIÓN DE LOS DATOS DEL EJEMPLO DEL RENDIMIENTO EN MILLAS POR GALÓN

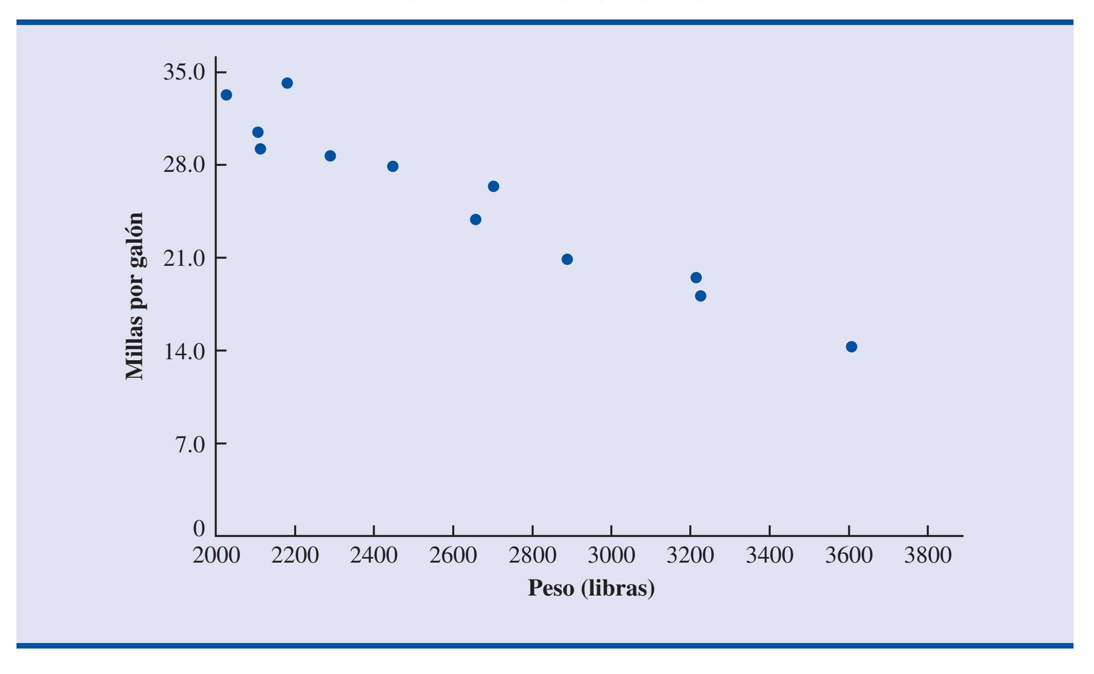


FIGURA 16.9 RESULTADOS DE MINITAB PARA EL EJEMPLO DEL RENDIMIENTO EN MILLAS POR GALÓN

```
The regression equation is
MPG = 56.1 - 0.0116 Weight

Predictor          Coef     SE Coef        T      p
Constant         56.096       2.582    21.72  0.000
Weight       -0.0116436   0.0009677   -12.03  0.000

S = 1.671   R-sq = 93.5%   R-sq(adj) = 92.9%

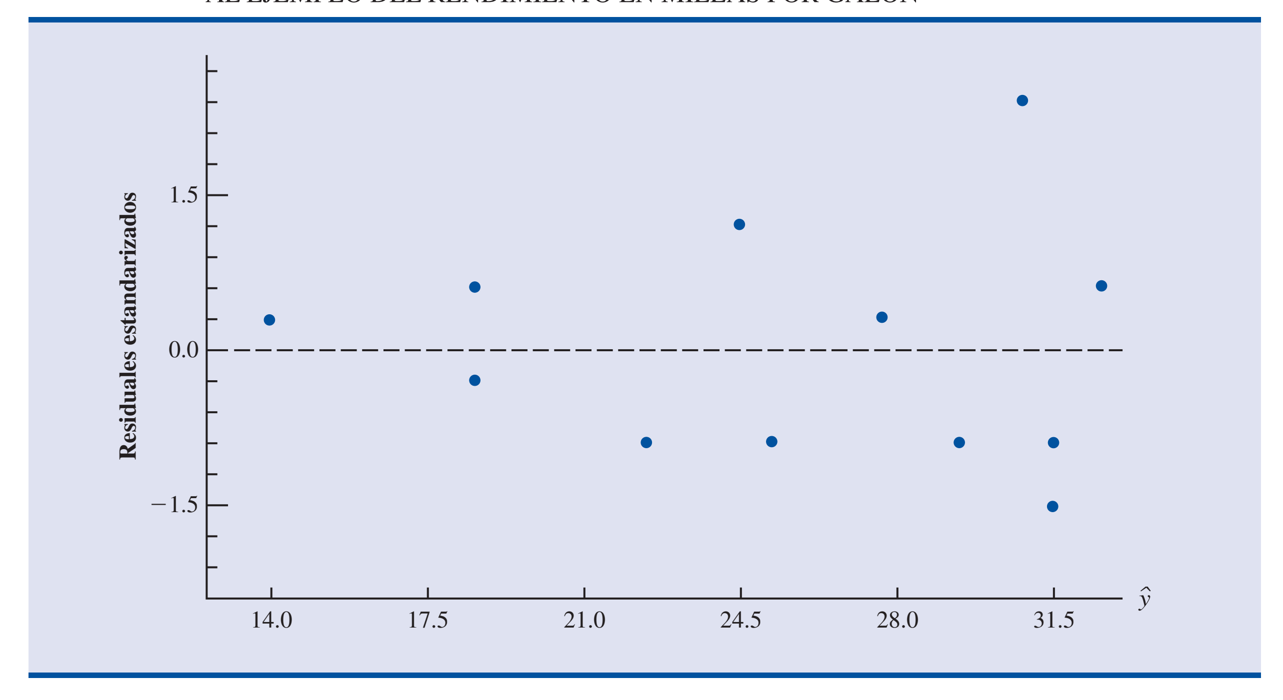
Analysis of Variance

SOURCE          DF      SS      MS       F      p
Regression       1  403.98  403.98  144.76  0.000
Residual Error  10   27.91    2.79
Total           11  431.88

Unusual Observations
Obs  Weight     MPG     Fit  SE Fit  Residual  St Resid
  3    2180  34.200  30.713   0.644     3.487     2.26R

R denotes an observation with a large standardized residual.
```

FIGURA 16.10 GRÁFICA DE LOS RESIDUALES ESTANDARIZADOS CORRESPONDIENTES AL EJEMPLO DEL RENDIMIENTO EN MILLAS POR GALÓN



garitmos base e = 2.71828... (logaritmos naturales). Aquí se empleará la transformación a logaritmos naturales de los datos de millas por galón y se desarrollará la ecuación estimada de regresión que relaciona el peso con el logaritmo natural de las millas por galón. En la figura 16.11 se muestra la ecuación de regresión que se obtiene al emplear el logaritmo natural de millas por galón como variable dependiente; esta ecuación aparece rotulada como LogeMPG. En la figura 16.12 se presenta la gráfica de los correspondientes residuales estandarizados.

Al observar la gráfica de residuales de la figura 16.12, se ve que la forma de cuña ha desaparecido. Además, ninguna de las observaciones ha sido identificada como una observación cuyo

FIGURA 16.11 RESULTADOS DE MINITAB PARA EL EJEMPLO DEL RENDIMIENTO EN MILLAS POR GALÓN: TRANSFORMACIÓN LOGARÍTMICA

```
The regression equation is
LogeMPG = 4.52 -0.000501 Weight

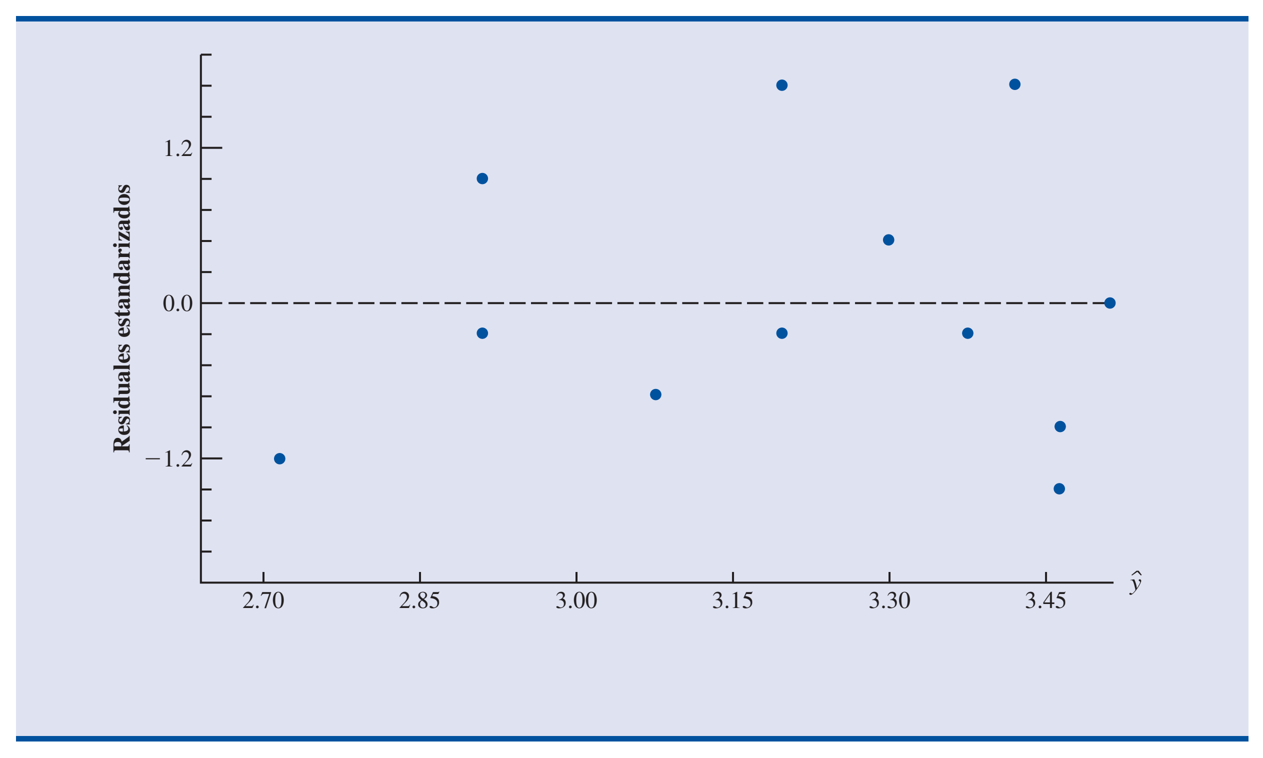
Predictor          Coef     SE Coef        T      p
Constant        4.52423     0.09932    45.55  0.000
Weight      -0.00050110  0.00003722   -13.46  0.000

S = 0.06425   R-sq = 94.8%   R-sq(adj) = 94.2%

Analysis of Variance

SOURCE          DF       SS         MS       F      p
Regression       1  0.74822    0.74822  181.22  0.000
Residual Error  10  0.04129    0.00413
Total           11  0.78950
```

FIGURA 16.12 GRÁFICA DE RESIDUALES ESTANDARIZADOS DEL EJEMPLO DEL RENDIMIENTO EN MILLAS POR GALÓN: TRANSFORMACIÓN LOGARÍTMICA



residual estandarizado sea grande. El modelo en el que se emplea como variable dependiente el logaritmo de las millas por galón es estadísticamente significativo y proporciona un ajuste excelente a los datos observados. Por tanto, se recomendará usar la ecuación estimada de regresión

$$\text{LogeMPG} = 4.52 - 0.000501 \text{ Weight (Peso)}$$

Para estimar el rendimiento en millas por galón de un automóvil que pese 2 500 libras, se obtiene primero una estimación del logaritmo del rendimiento de millas por galón.

$$\text{LogeMPG} = 4.52 - 0.000501(2500) = 3.2675$$

La estimación de las millas por galón se obtiene al hallar el número cuyo logaritmo natural es 3.2675. Al emplear una calculadora con función exponencial o elevar e a la potencia 3. 2675, se obtienen 26.2 millas por galón.

Otro método para problemas con varianza no constante es usar como variable dependiente $1/y$, en lugar de y. A este tipo de transformación se le llama *transformación recíproca.* Por ejemplo, si la variable dependiente se mide en millas por galón, la transformación recíproca dará como resultado una nueva variable dependiente cuyas unidades serán 1/(millas por galón) o galones por milla. No hay manera de determinar qué transformación funcionará mejor, si una transformación logarítmica o una transformación recíproca, si no es probándolas.

Modelos no lineales que son intrínsecamente lineales

A los modelos que tienen parámetros ($\beta_0, \beta_1, \ldots, \beta_p$) con exponentes distintos a 1 se les conoce como modelos no lineales. Sin embargo, en el caso de modelos exponenciales, se puede realizar una transformación de las variables que permita realizar el análisis de regresión mediante la

ecuación (16.1), el modelo lineal general. En el modelo general exponencial se tiene la siguiente ecuación de regresión

$$E(y) = \beta_0 \beta_1^x \tag{16.7}$$

Este modelo es adecuado cuando la variable dependiente y aumenta o disminuye en un porcentaje constante, en lugar de en una cantidad fija, a medida que x aumenta.

Por ejemplo, suponga que las ventas y de un producto se relacionan con los gastos en publicidad x (en miles de dólares) de acuerdo con el siguiente modelo exponencial.

$$E(y) = 500(1.2)^x$$

Por tanto, para $x = 1$, $E(y) = 500(1.2)^1 = 600$; para $x = 2$, $E(y) = 500(1.2)^2 = 720$; para $x = 3$, $E(y) = 500(1.2)^3 = 864$. Observe, que en este caso, $E(y)$ no aumenta en una cantidad constante, sino en un porcentaje constante; el porcentaje de aumento es 20%.

Al sacar logaritmos a ambos lados de la ecuación (16.7) se puede transformar este modelo no lineal en un modelo lineal.

$$\log E(y) = \log \beta_0 + x \log \beta_1 \tag{16.8}$$

Ahora, si $y' = \log E(y)$, $\beta_0' = \log \beta_0$, y $\beta_1' = \log \beta_1$, la ecuación (16.8) se expresa como

$$y' = \beta_0' + \beta_1' x$$

Ahora es claro que se puede emplear la fórmula de la regresión lineal simple para obtener estimaciones de β_0' y de β_1'. Al denotar estas estimaciones como b_0' y b_1' se obtiene la siguiente ecuación estimada de regresión.

$$\hat{y}' = b_0' + b_1' x \tag{16.9}$$

Para obtener predicciones para la variable dependiente original y dado un valor de x, primero se sustituye el valor de x en la ecuación (16.9) y se calcula $\hat{y}'$. El antilogaritmo de $\hat{y}'$ será la predicción de y o el valor esperado de y.

Muchos modelos no lineales pueden ser transformados a un modelo lineal equivalente. Sin embargo, tales modelos han tenido pocas aplicaciones en el comercio y la economía. Además, los fundamentos matemáticos para el estudio de tales modelos quedan fuera del alcance de este libro.

Ejercicios

Métodos



1. Considere los datos siguientes para las variables x y y.

x	22	24	26	30	35	40
y	12	21	33	35	40	36



a. Con estos datos obtenga una ecuación estimada de regresión de la forma $\hat{y} = b_0 + b_1 x$.

b. Use los resultados del inciso a para probar si existe una relación significativa entre x y y. Use $\alpha = 0.05$

c. Obtenga el diagrama de dispersión de estos datos. ¿Este diagrama de dispersión sugiere una ecuación estimada de regresión de la forma $\hat{y} = b_0 + b_1 x + b_2 x^2$? Explique.

d. Con estos datos obtenga una ecuación estimada de regresión de la forma $\hat{y} = b_0 + b_1x + b_2x^2$

e. Remítase al inciso d. ¿La relación entre x, x^2 y y es significativa? Use $\alpha = 0.05$

f. Prediga el valor de y para $x = 25$.

2. Considere los datos siguientes para las variables x y y.

x	9	32	18	15	26
y	10	20	21	16	22

a. Con estos datos obtenga una ecuación estimada de regresión de la forma $\hat{y} = b_0 + b_1x$. Presente un comentario sobre lo adecuado de esta ecuación para predecir y.

b. Con estos datos obtenga una ecuación estimada de regresión de la forma $\hat{y} = b_0 + b_1x + b_2x^2$. Dé un comentario sobre lo adecuado de esta relación para predecir y.

c. Prediga el valor de y para $x = 20$.

3. Considere los datos siguientes para las variables x y y.

x	2	3	4	5	7	7	7	8	9
y	4	5	4	6	4	6	9	5	11

a. ¿Parece haber una relación lineal entre x y y? Explique.

b. Obtenga la ecuación estimada de regresión que relaciona x y y.

c. Dada la ecuación estimada de regresión obtenida en el inciso b, grafique los residuales estandarizados contra $\hat{y}$. ¿Las suposiciones del modelo parecen satisfacerse? Explique.

d. Realice una transformación logarítmica de la variable dependiente y. Obtenga una ecuación estimada de regresión, emplee la variable dependiente transformada. ¿Las suposiciones del modelo parecen satisfacerse cuando se usa una variable dependiente transformada? En este caso, ¿la transformación recíproca funciona mejor? Explique.

Aplicaciones

4. El departamento de autopistas estudia la relación entre flujo de tráfico y velocidad. Se considera que el modelo siguiente es el adecuado.

$$y = \beta_0 + \beta_1 x + \epsilon$$

donde

y = flujo de tráfico en vehículos por hora
x = velocidad de los vehículos en millas por hora

Los siguientes datos fueron recolectados durante "horas pico" en las seis principales autopistas que salen de la ciudad.

Flujo de tráfico (y)	**Velocidad de los vehículos (x)**
1256	35
1329	40
1226	30
1335	45
1349	50
1124	25

a. Obtenga con estos datos una ecuación estimada de regresión.

b. Use $\alpha = 0.01$ para probar la significancia de la relación.



5. Para continuar con el problema del ejercicio 4, se sugiere emplear la siguiente ecuación estimada de regresión curvilínea

$$\hat{y} = b_0 + b_1x + b_2x^2$$

a. Use los datos del problema 4 para estimar los parámetros de esta ecuación estimada de regresión.
b. Use $\alpha = 0.01$ para probar la significancia de la relación.
c. Estime el flujo de tráfico en vehículos por hora correspondiente a 38 millas por hora.

6. En un estudio sobre instalaciones para servicios de emergencia se investigó la relación entre el número de instalaciones y la distancia promedio a recorrer para dar el servicio de emergencia. En la tabla siguiente se presentan los datos obtenidos.

Número de instalaciones	Distancia promedio (millas)
9	1.66
11	1.12
16	0.83
21	0.62
27	0.51
30	0.47

a. Trace el diagrama de dispersión de estos datos, considere la distancia promedio a recorrer como la variable dependiente.
b. ¿Un modelo lineal simple será apropiado? Explique.
c. Con estos datos obtenga la ecuación estimada de regresión que mejor explica la relación entre las dos variables.

7. Un factor importante al comprar un monitor para computadora es el campo de visión. Un monitor que tenga un campo de visión amplio permite tener una imagen aceptable con sólo girar ligeramente la cabeza y una persona de pie cerca del monitor logra ver claramente la imagen de la pantalla. Después de una revisión de monitores LCD de 19 pulgadas, *PC World* encontró que aunque todos los monitores probados aseguraban arcos de 170 grados —tanto horizontal como verticalmente— el rango real de los monitores iba de 108 a 167 grados. En la tabla siguiente se da el ángulo de visión horizontal de ocho monitores de 19 pulgadas y la evaluación dada por *PC World* con base en la calidad de la imagen, el precio y en las políticas de soporte técnico (*PC World*, febrero de 2003).



Monitors

Monitor	Ángulo	Evaluación
Samsung SyncMaster 191T	167	86
ViewSonic VX900	159	82
Sceptre Technologies X9S-Naga	126	81
Planar PL191M	108	81
Dell UltraSharp 1900FP	153	81
AOC LM914	123	81
KDS USA Radius Rad-9	118	80
NEC MultiSync LCD 1920NX	123	80
Iiyama Pro Lite 4821DT-BK	119	80

a. Desarrolle un diagrama de dispersión de estos datos, emplee como variable independiente el ángulo de visión horizontal.
b. ¿Un modelo de regresión lineal simple es apropiado?
c. Obtenga una ecuación estimada de regresión que explique la relación entre estas dos variables.

8. Corvette, Ferrari y Jaguar fabricaron varios automóviles clásicos con un valor que aún sigue en aumento. En la tabla siguiente, basada en el sistema Martin de evaluación de automóviles de colección, se presenta la evaluación de su rareza (1-20) y el precio ($ miles) de 15 automóviles clásicos (www.businessweek.com, febrero de 2006).



ClassicCars

Año	Fabricante	Modelo	Evaluación	Precio ($ miles)
1984	Chevrolet	Corvette	18	1600
1956	Chevrolet	Corvette 265/225-hp	19	4000
1963	Chevrolet	Corvette coupe (340-bhp 4-speed)	18	1000
1978	Chevrolet	Corvette coupe Silver Anniversary	19	1300
1960-1963	Ferrari	250 GTE 2+2	16	350
1962-1964	Ferrari	250 GTL Lusso	19	2650
1962	Ferrari	250 GTO	18	375
1967-1968	Ferrari	275 GTB/4 NART Spyder	17	450
1968-1973	Ferrari	365 GTB/4 Daytona	17	140
1962-1967	Jaguar	E-type OTS	15	77.5
1969-1971	Jaguar	E-type Series II OTS	14	62
1971-1974	Jaguar	E-type Series III OTS	16	125
1951-1954	Jaguar	XK 120 roadster (steel)	17	400
1950-1953	Jaguar	XK C-type	16	250
1956-1957	Jaguar	XKSS	13	70

a. Dé el diagrama de dispersión de estos datos, emplee la evaluación de la rareza como variable independiente y el precio como variable independiente. ¿Es apropiado un modelo de regresión lineal simple?
b. Obtenga una ecuación estimada de regresión simple en la cual las variables independientes sean x = evaluación de la rareza y x^2.
c. Considere la relación no lineal dada por la ecuación (16.7). Use logaritmos para obtener una ecuación estimada de regresión para este modelo.
d. ¿Cuál de las ecuaciones estimadas de regresión prefiere, la del inciso b o la del inciso c? Explique.

9. Casi en todo el sistema de trenes ligeros de Estados Unidos se usan vagones eléctricos que corren sobre vías construidas a nivel de la calle. De acuerdo con la Administración de Tránsito Federal de Estados Unidos, el tren ligero es uno de los medios de transporte más seguros, con una tasa de accidentes de 0.99 accidentes por cada millón de millas de pasajeros en comparación con 2.29 en los autobuses. En la tabla siguiente se presenta, para algunos de los sistemas de tren ligero de Estados Unidos, el número de millas de vía y el número de pasajeros, en miles, que utilizan el transporte público en un día entre semana.



LightRail

Ciudad	Millas	Pasajeros
Los Ángeles	22	70
San Diego	47	75
Portland	38	81
Sacramento	21	31
San Jose	31	30
San Francisco	73	164
Philadelphia	69	84
Boston	51	231
Denver	17	35
Salt Lake City	18	28

(*continúa*)

Ciudad	Millas	Pasajeros
Dallas	44	39
Nueva Orleans	16	14
San Luis	34	42
Pittsburgh	18	25
Buffalo	6	23
Cleveland	15	15
Newark	9	8

a. Dé el diagrama de dispersión de estos datos, emplee como variable independiente el número de millas de vía. ¿Es apropiado emplear un modelo de regresión lineal?
b. Use un modelo de regresión lineal simple para obtener una ecuación estimada de regresión que sirva para predecir el número de pasajeros por día entre semana, dado que conoce el número de millas de vía. Construya una gráfica de residuales estandarizados. Con base en la gráfica de residuales estandarizados diga si el modelo de regresión lineal simple es apropiado.
c. Realice una transformación logarítmica de la variable dependiente. Obtenga una ecuación estimada de regresión, emplee la variable dependiente transformada. ¿Las suposiciones del modelo al usar la variable dependiente transformada se satisfacen?
d. Realice una transformación recíproca de la variable dependiente. Obtenga una ecuación estimada de regresión, emplee la variable dependiente transformada.
e. ¿Cuál de las ecuaciones de regresión estimada recomienda? Explique.



Determinación de cuándo agregar o quitar variables

En esta sección se mostrará el uso de la prueba F para determinar si es ventajoso agregar una o más variables independientes a un modelo de regresión múltiple. Esta prueba se basa en determinar la disminución del valor de la suma de cuadrados debidos al error al agregar una o más variables independientes al modelo. Primero se ilustrará el uso del modelo en el contexto del ejemplo de Butler Trucking.

En el capítulo 15 se presentó el modelo de Butler Trucking para ilustrar el uso del análisis de regresión múltiple. Como recordará, los directivos de esta empresa deseaban obtener una ecuación estimada de regresión para predecir el tiempo total del recorrido diario de sus camiones repartidores, a partir de dos variables independientes: millas recorridas y número de entregas. Al usar como única variable independiente el número de millas recorridas x_1, se obtuvo la siguiente ecuación estimada de regresión, mediante el método de mínimos cuadrados.

$$\hat{y} = 1.27 + 0.0678x_1$$

En el capítulo 15 se mostró que la suma de cuadrados debidos al error con este modelo era: SCE = 8.029. Cuando se agregó al modelo otra variable independiente, número de entregas x_2, se obtuvo la siguiente ecuación estimada de regresión.

$$\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$$

La suma de cuadrados debidos al error en este modelo fue SCE = 2.299. Agregar x_2 dio como resultado una reducción en el valor de SCE. La pregunta es: ¿La adición de la variable x_2 conduce a una reducción *significativa* en el valor de la SCE?

Se empleará la notación siguiente: SCE(x_1) para denotar la suma de cuadrados debidos al error cuando x_1 es la única variable independiente del modelo; SCE(x_1, x_2) para denotar la suma

de cuadrados debidos al error, cuando tanto x_1 como x_2 son las variables del modelo, y así sucesivamente. Por tanto, la disminución del valor de la SCE que se obtuvo al adicionar x_2 al modelo que sólo tenía como variable independiente a x_1 es

$$\text{SCE}(x_1) - \text{SCE}(x_1, x_2) = 8.029 - 2.299 = 5.730$$

Para determinar si esta reducción es significativa se realiza una prueba F.

El numerador del estadístico F es la disminución en el valor de SCE dividida entre la cantidad de variables independientes agregadas al modelo original. En este caso, sólo se agregó una variable, x_2, por lo que el numerador del estadístico F es

$$\frac{\text{SCE}(x_1) - \text{SCE}(x_1, x_2)}{1} = 5.730$$

Lo que se obtiene es una medida de la disminución de SCE por variable independiente añadida al modelo. El denominador del estadístico F es el cuadrado medio debido al error en el modelo que contiene todas las variables independientes. En el caso del ejemplo de Butler Trucking, esto corresponde al modelo que tiene tanto a x_1 como a x_2; por tanto $p = 2$ y

$$\text{CME} = \frac{\text{SCE}\ (x_1, x_2)}{n - p - 1} = \frac{2.299}{7} = 0.3284$$

El siguiente estadístico F es la base para probar si la adición de x_2 es estadísticamente significativa.

$$F = \frac{\dfrac{\text{SCE}(x_1) - \text{SCE}(x_1, x_2)}{1}}{\dfrac{\text{SCE}(x_1, x_2)}{n - p - 1}} \tag{16.10}$$

El número de grados de libertad en el numerador de este estadístico F es igual al número de variables agregadas al modelo y el número de grados en el denominador es igual a $n - p - 1$.

En el caso del problema de Butler Trucking se obtiene

$$F = \frac{\dfrac{5.730}{1}}{\dfrac{2.299}{7}} = \frac{5.730}{0.3284} = 17.45$$

Si consulta la tabla 4 del apéndice B, se encuentra que para el nivel de significancia 0.05, $F_{0.05} = 5.59$, por lo que se puede rechazar la hipótesis de que x_2 no sea estadísticamente significativa; en otras palabras, al agregar x_2 al modelo en el que sólo se tiene como variable independiente x_1, se obtiene una disminución significativa en la suma de los cuadrados debido al error.

Cuando se desea probar la significancia de agregar sólo una variable independiente al modelo, el resultado que se obtiene con la prueba F que se acaba de describir, también se obtiene con la prueba t para la significancia de uno solo de los parámetros (descrita en la sección 15.4). El estadístico F que se acaba de calcular es el cuadrado del estadístico t que se usa para probar la significancia de un solo parámetro.

Puesto que, cuando se agrega una sola variable independiente al modelo, la prueba t es equivalente a la prueba F, esto permite aclarar el uso adecuado de la prueba t para probar la significancia de uno de los parámetros. Si uno de los parámetros no es significativo, la variable correspondiente puede ser eliminada del modelo. Pero, si la prueba t indica que hay dos o más

parámetros que no son significativos, nunca se debe eliminar del modelo más de una variable independiente, con base en la prueba t; cuando se elimina una variable, puede resultar que una segunda variable, que inicialmente no era significativa, se vuelva significativa.

Ahora cabe considerar si la adición de más de una variable independiente –como conjunto– da como resultado que haya una reducción significativa de la suma de los cuadrados debidos al error.

Caso general

Considere el siguiente modelo de regresión múltiple en el que intervienen q variables independientes, donde $q < p$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \epsilon \tag{16.11}$$

Si a este modelo se le agregan las variables $x_{q+1}, x_{q+2}, \ldots, x_p$, se obtiene un modelo con p variables independientes.

$$\begin{aligned} y = {} & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q \\ & + \beta_{q+1} x_{q+1} + \beta_{q+2} x_{q+2} + \cdots + \beta_p x_p + \epsilon \end{aligned} \tag{16.12}$$

Para probar si la adición de $x_{q+1}, x_{q+2}, \ldots, x_p$, es estadísticamente significativa, las hipótesis nula y alternativa pueden plantearse como sigue.

H_0: $\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$
H_a: Uno o más de los parámetros no es igual a cero

El siguiente estadístico F aporta la base para probar si la adición de estas variables independientes es estadísticamente significativa.

$$F = \frac{\dfrac{\text{SCE}(x_1, x_2, \ldots, x_q) - \text{SCE}(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{p - q}}{\dfrac{\text{SCE}(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{n - p - 1}} \tag{16.13}$$

Este valor F calculado se compara con F_α, el valor en la tabla para $p - q$ grados de libertad en el numerador y $n - p - 1$ grados de libertad en el denominador. Si $F > F_\alpha$ se rechaza H_0 y se concluye que el conjunto de variables independientes agregadas es estadísticamente significativo. Observe que si $q = 1$ y $p = 2$, la ecuación (16.13) se reduce a la ecuación (16.10).

Muchos paquetes de software, como Minitab, proporcionan sumas de cuadrados que corresponden al orden en el que cada variable independiente entra al modelo; en tales casos se simplifican los cálculos de la prueba F para determinar si agregar o eliminar un conjunto de variables.

Para muchos estudiantes encontrar la ecuación (16.13) resulta un poco complicado. Para dar una descripción un poco más sencilla de este cociente F, al modelo que tiene la menor cantidad de variables independientes se le denomina modelo reducido y al modelo que tiene la mayor cantidad de variables independientes se le denomina modelo completo. Si SCE(reducido) denota la suma de los cuadrados debidos al error del modelo reducido y SCE(completo) denota la suma de los cuadrados debidos al error del modelo completo, el numerador de la ecuación (16.13) se expresa como

$$\frac{\text{SCE(reducido)} - \text{SCE(completo)}}{\text{número de términos extra}} \tag{16.14}$$

Observe que "número términos extra" denota la diferencia entre el número de variables independientes en el modelo completo y el número de variables independientes en el modelo reducido. El denominador de la ecuación (16.13) es la suma de los cuadrados debidos al error en el modelo completo dividida entre los correspondientes grados de libertad; en otras palabras, el denomi-

nador es el cuadrado medio debido al error en el modelo completo. Si se denota el cuadrado medio debido al error del modelo completo como CME(completo) se puede escribir

$$F = \frac{\dfrac{\text{SCE(reducido)} - \text{SCE(completo)}}{\text{número de términos extra}}}{\text{CME(completo)}} \tag{16.15}$$

Para ilustrar el uso de este estadístico *F*, suponga que se tiene un problema de regresión que tiene 30 observaciones. En un modelo en el que intervienen las variables independientes x_1, x_2 y x_3 la suma de los cuadrados debida al error es 150 y en un segundo modelo en el que las variables independientes son x_1, x_2, x_3, x_4 y x_5, la suma de los cuadrados debida al error es 100. ¿La adición de las variables x_4 y x_5 produjo una reducción significativa de la suma de los cuadrados debida al error?

Observe, primero, que el número de grados de libertad para STC es $30 - 1 = 29$ y que el número de grados de libertad para la suma de cuadrados debida a la regresión para el modelo completo es cinco (el número de variables independientes en el modelo completo). Por tanto, los grados de libertad para la suma de los cuadrados debida al error en el modelo completo es $29 - 5 = 24$, entonces CME(completo) $= 100/24 = 4.17$. Así, el estadístico *F* es

$$F = \frac{\dfrac{150 - 100}{2}}{4.17} = 6.00$$

Este valor *F* que se ha calculado se compara con el valor *F* que se encuentra en la tabla para dos grados de libertad en el numerador y 24 grados de libertad en el denominador. Para el nivel de significancia 0.05, en la tabla 4 del apéndice B se encuentra $F_{0.05} = 3.40$. Como $F = 6.00$ es mayor que 3.40, se concluye que la adición de las variables x_4 y x_5 es estadísticamente significativa.

Uso del valor-*p*

También puede usarse el criterio del valor-*p* para determinar si resulta ventajoso agregar una o más variables independientes a un modelo de regresión múltiple. En el ejemplo anterior se mostró cómo realizar la prueba *F* para determinar si la adición de dos variables independientes, x_4 y x_5, a un modelo con tres variables independientes, x_1, x_2 y x_3, era estadísticamente significativo. En ese ejemplo el valor que se obtuvo para el estadístico *F* fue 6.00 y se concluyó (por comparación de $F = 6.00$ con el valor crítico $F_{0.05} = 3.40$) que la adición de las variables x_4 y x_5 era significativa. El valor-*p* que corresponde a $F = 6.00$ (2 grados de libertad en el numerador y 24 grados de libertad en el denominador) es 0.008. Como el valor-$p = 0.008 < \alpha = 0.05$, también se concluye que la adición de las dos variables independientes es significativa. Al emplear las tablas de la distribución *F* es difícil determinar directamente el valor-*p*, pero los paquetes de software como Minitab o Excel facilitan este cálculo.

NOTAS Y COMENTARIOS

El cálculo del estadístico *F* también se basa en las sumas de cuadrados debida a la regresión. Para mostrar esta forma del estadístico *F*, se nota primero que

$$\text{SCE(reducido)} = \text{STC} - \text{SCR(reducido)}$$
$$\text{SCE(completo)} = \text{STC} - \text{SCR(completo)}$$

Por tanto

$$\begin{aligned}\text{SCE(reducido)} - \text{SCE(completo)} &= [\text{STC} - \text{SCR(reducido)}] - [\text{STC} - \text{SCR(completo)}] \\ &= \text{SCR(completo)} - \text{SCR(reducido)}\end{aligned}$$

Así,

$$F = \frac{\dfrac{\text{SCR(completo)} - \text{SCR(reducido)}}{\text{número de términos extra}}}{\text{CME(completo)}}$$

Ejercicios

Métodos

10. En un análisis de regresión en el que se emplearon 27 observaciones, se obtuvo la siguiente ecuación estimada de regresión.

$$\hat{y} = 25.2 + 5.5x_1$$

Para esta ecuación estimada de regresión STC = 1 550 y SCE = 520.

a. Utilice $\alpha = 0.05$ y pruebe si x_1 es significativa.

Suponga que a este modelo le agrega las variables x_2 y x_3 y obtiene la ecuación de regresión siguiente.

$$\hat{y} = 16.3 + 2.3x_1 + 12.1x_2 - 5.8x_3$$

Para esta ecuación estimada de regresión STC = 1 550 y SCE = 100.

b. Use una prueba F y 0.05 como nivel de significancia para determinar si x_2 y x_3 contribuyen significativamente al modelo.



11. En un análisis de regresión en el que se emplearon 30 observaciones, se obtuvo la siguiente ecuación estimada de regresión.

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

Para esta ecuación estimada de regresión STC = 1 805 y SCR = 1 760.

a. Con $\alpha = 0.05$, pruebe la significancia de la relación entre las variables.

Suponga que de este modelo elimina las variables x_1 y x_4 y obtiene la siguiente ecuación estimada de regresión.

$$\hat{y} = 11.1 - 3.6x_2 + 8.1x_3$$

Para esta ecuación estimada de regresión STC = 1 805 y SCR = 1 705.

b. Calcule SCE(x_1, x_2, x_3 y x_4)
c. Calcule SCE(x_2, x_3)
d. Use una prueba F y 0.05 como nivel de significancia para determinar si x_1 y x_4 contribuyen significativamente al modelo.

Aplicaciones



12. La Ladies Professional Golfers Association (LPGA, por sus siglas en inglés) lleva estadísticas sobre el desempeño y las ganancias de sus miembros en la LPGA Tour. En el archivo titulado LPGA del disco compacto se presentan las estadísticas de fin de año sobre el desempeño de las 30 jugadoras que obtuvieron las mayores ganancias en la LPGA Tour de 2005 (www.lpga.com, 2006). Earnings (ganancias) ($ miles) son los ingresos totales en miles de dólares; Scoring Avg., es la puntuación promedio de una jugadora en todos los eventos; Greens in Reg., es el porcentaje de las veces que una jugadora llega al green en regulación; Putting Avg., es el promedio de putts hechos en el green en regulación, y Sand Saves es el porcentaje de veces que la jugadora

logra “subir y bajar” (“up and down”) cuando se encuentra en un búnker de arena al lado del green.

a. Obtenga una ecuación estimada de regresión que sirva para predecir Scoring Avg. dado Greens in Reg.
b. Obtenga una ecuación estimada de regresión que sirva para predecir Scoring Avg. dados Greens in Reg., Putting Avg. y Sand Saves.
c. Con un nivel de significancia 0.05 pruebe si las dos variables independientes agregadas en el inciso b, Putting Avg. y Sand Saves, contribuyen significativamente a la ecuación estimada de regresión obtenida en el inciso a. Explique.



LPGATour

13. Vaya al ejercicio 12.
a. Obtenga una ecuación estimada de regresión que sirva para predecir Earnings, conociendo Putting Avg.
b. Obtenga una ecuación estimada de regresión que sirva para predecir Earnings, conociendo Putting Avg. y Sand Saves.
c. Emplee como nivel de significancia 0.05 y pruebe si las dos variables independientes agregadas en el inciso b, Putting Avg. y Sand Saves, contribuyen significativamente a la ecuación estimada de regresión obtenida en el inciso a. Explique.
d. En general, puntuaciones más bajas llevan a ganancias más altas. Para investigar esta opción para predecir las Earnings, obtenga una ecuación estimada de regresión que sirva para predecir Earnings, Scoring Avg. ¿Preferiría emplear esta ecuación para predecir las ganancias (Earnings) o la ecuación obtenida en el inciso b? Explique.

14. En un estudio realizado a lo largo de 10 años por la American Heart Association se obtuvieron datos acerca de la relación entre edad, presión sanguínea y fumar con el riesgo a sufrir un infarto. Los datos que se presentan a continuación son parte de este estudio. El riesgo se interpreta como la probabilidad (multiplicada por 100) de que el paciente sufra un infarto en los próximos 10 años. Para la variable fumar, defina una variable ficticia que tome el valor 1 si la persona es fumadora y el valor 0 si no es fumadora.



Stroke

Riesgo	Edad	Presión	Fumador
12	57	152	0
24	67	163	0
13	58	155	0
56	86	177	1
28	59	196	0
51	76	189	1
18	56	155	1
31	78	120	0
37	80	135	1
15	78	98	0
22	71	152	0
36	70	173	1

(*continúa*)

Riesgo	Edad	Presión	Fumador
15	67	135	1
48	77	209	1
15	60	199	0
36	82	119	1
8	66	166	0
34	80	125	1
3	62	117	0
37	59	207	1

a. Obtenga una ecuación estimada de regresión que sirva para predecir el riesgo de sufrir un infarto, dados edad y presión sanguínea.
b. Considere la adición de dos variables independientes al modelo obtenido en el inciso a, una para la interacción entre edad y presión sanguínea y otra que indique si la persona es o no un fumador. Obtenga una ecuación estimada de regresión, emplee estas cuatro variables independientes.
c. Emplee como nivel de significancia 0.05, realice una prueba para determinar si la adición de la variable de la interacción y la variable fumador contribuyen significantemente a la ecuación estimada de regresión obtenida en el inciso a.

15. La National Football League, NFL, evalúa a sus prospectos con una escala que va del 5 al 9. Estas evaluaciones se interpretan como sigue: 8-9 deberá empezar el año próximo; 7.0-7.9 deberá empezar; 6.0-6.9 servirán de respaldo al equipo y 5.0-5.9 pueden formar parte del club y contribuir. En la tabla siguiente se da posición, peso, tiempo en segundos para correr 40 yardas y la evaluación dada por la NFL a 40 prospectos (*USA Today*, 14 de abril de 2000).

archivo en CD

NFL

Observación	Nombre	Posición	Peso	Tiempo	Evaluación
1	Peter Warrick	Receptor	194	4.53	9.0
2	Plaxico Burress	Receptor	231	4.52	8.8
3	Sylvester Morris	Receptor	216	4.59	8.3
4	Travis Taylor	Receptor	199	4.36	8.1
5	Laveranues Coles	Receptor	192	4.29	8.0
6	Dez White	Receptor	218	4.49	7.9
7	Jerry Porter	Receptor	221	4.55	7.4
8	Ron Dugans	Receptor	206	4.47	7.1
9	Todd Pinkston	Receptor	169	4.37	7.0
10	Dennis Northcutt	Receptor	175	4.43	7.0
11	Anthony Lucas	Receptor	194	4.51	6.9
12	Darrell Jackson	Receptor	197	4.56	6.6
13	Danny Farmer	Receptor	217	4.60	6.5
14	Sherrod Gideon	Receptor	173	4.57	6.4
15	Trevor Gaylor	Receptor	199	4.57	6.2
16	Cosey Coleman	Guardia	322	5.38	7.4
17	Travis Claridge	Guardia	303	5.18	7.0
18	Kaulana Noa	Guardia	317	5.34	6.8
19	Leander Jordan	Guardia	330	5.46	6.7
20	Chad Clifton	Guardia	334	5.18	6.3
21	Manula Savea	Guardia	308	5.32	6.1
22	Ryan Johanningmeir	Guardia	310	5.28	6.0
23	Mark Tauscher	Guardia	318	5.37	6.0
24	Blaine Saipaia	Guardia	321	5.25	6.0
25	Richard Mercier	Guardia	295	5.34	5.8
26	Damion McIntosh	Guardia	328	5.31	5.3
27	Jeno James	Guardia	320	5.64	5.0

Observación	Nombre	Posición	Peso	Tiempo	Evaluación
28	Al Jackson	Guardia	304	5.20	5.0
29	Chris Samuels	Tackle ofensivo	325	4.95	8.5
30	Stockar McDougle	Tackle ofensivo	361	5.50	8.0
31	Chris McIngosh	Tackle ofensivo	315	5.39	7.8
32	Adrian Klemm	Tackle ofensivo	307	4.98	7.6
33	Todd Wade	Tackle ofensivo	326	5.20	7.3
34	Marvel Smith	Tackle ofensivo	320	5.36	7.1
35	Michael Thompson	Tackle ofensivo	287	5.05	6.8
36	Bobby Williams	Tackle ofensivo	332	5.26	6.8
37	Darnell Alford	Tackle ofensivo	334	5.55	6.4
38	Terrance Beadles	Tackle ofensivo	312	5.15	6.3
39	Tutan Reyes	Tackle ofensivo	299	5.35	6.1
40	Greg Robinson-Ran	Tackle ofensivo	333	5.59	6.0

a. Dé una variable ficticia para la posición de los jugadores.
b. Obtenga una ecuación estimada de regresión que indique la relación entre evaluación y posición, peso y tiempo para correr 40 yardas.
c. Emplee 0.05 como nivel de significancia, pruebe si la ecuación estimada de regresión obtenida en el inciso b representa una relación significante entre las variables independientes y la variable dependiente.
d. ¿La posición es un factor significativo para la evaluación del jugador? Use $\alpha = 0.05$. Explique.



16.3 Análisis de un problema mayor

Al introducir el análisis de regresión múltiple, se usó ampliamente el ejemplo de Butler Trucking. Al explorar los conceptos fue una ventaja que este problema fuera pequeño. Sin embargo, este problema pequeño dificulta ilustrar algunas de las cuestiones relacionadas con la selección de variables que intervienen en la construcción de un modelo. Para dar un ejemplo de los procesos de selección de variables que se estudian en la sección siguiente, se introduce un conjunto de datos que consta de 25 observaciones con ocho variables independientes. El doctor David W. Cravens del departamento de marketing de la Texas Christian University otorgó el permiso para emplear estos datos. Por esta razón a este conjunto de datos se le llamará datos de Cravens.*

Los datos de Cravens son de una empresa que tiene varios territorios de ventas, cada uno de los cuales le está asignado a un solo representante de ventas. Para determinar si diversas variables (independientes) predictoras podían explicar las ventas en cada uno de los territorios se realizó un análisis de regresión. A partir de una muestra de 25 territorios se obtuvieron los datos que se muestran en la tabla 16.5; en la tabla 16.6 se presenta la definición de las variables.

Como paso preliminar se considerarán los coeficientes de correlación entre cada par de variables. En la figura 16.13 se presenta la matriz de correlación obtenida con Minitab. Observe que el coeficiente de correlación muestral entre Sales y Time es 0.623, entre Sales y Poten es 0.598 y así sucesivamente.

Si observa los coeficientes de correlación entre las variables independientes, se dará cuenta de que la correlación entre Time y Accounts es 0.758; por tanto, si Accounts se usa como una de las variables independientes, Time no agregaría mucho poder explicatorio al modelo. Recuerde la prueba de la regla práctica que se vio en la sección 15.4, donde la multicolinealidad puede causar problemas si el valor absoluto del coeficiente de correlación muestral, entre cualesquiera dos de las variables independientes, es mayor que 0.7. Por tanto, siempre que sea posible,

*Para más detalles ver David W. Cravens, Robert B. Woodruff y Joe C. Stamper, "An Analytical Approach for Evaluating Sales Territory Performance", *Journal of Marketing*, 36 (enero, 1972): 31-37. Copyright © 1972 American Marketing Association.

TABLA 16.5 DATOS DE CRAVENS



Ventas	Antigüedad	Potencial	GastPubl	Participación	Cambio	Cuentas	Trabajo	Evaluación
3 669.88	43.10	74 065.1	4 582.9	2.51	0.34	74.86	15.05	4.9
3 473.95	108.13	58 117.3	5 539.8	5.51	0.15	107.32	19.97	5.1
2 295.10	13.82	21 118.5	2 950.4	10.91	−0.72	96.75	17.34	2.9
4 675.56	186.18	68 521.3	2 243.1	8.27	0.17	195.12	13.40	3.4
6 125.96	161.79	57 805.1	7 747.1	9.15	0.50	180.44	17.64	4.6
2 134.94	8.94	37 806.9	402.4	5.51	0.15	104.88	16.22	4.5
5 031.66	365.04	50 935.3	3 140.6	8.54	0.55	256.10	18.80	4.6
3 367.45	220.32	35 602.1	2 086.2	7.07	−0.49	126.83	19.86	2.3
6 519.45	127.64	46 176.8	8 846.2	12.54	1.24	203.25	17.42	4.9
4 876.37	105.69	42 053.2	5 673.1	8.85	0.31	119.51	21.41	2.8
2 468.27	57.72	36 829.7	2 761.8	5.38	0.37	116.26	16.32	3.1
2 533.31	23.58	33 612.7	1 991.8	5.43	−0.65	142.28	14.51	4.2
2 408.11	13.82	21 412.8	1 971.5	8.48	0.64	89.43	19.35	4.3
2 337.38	13.82	20 416.9	1 737.4	7.80	1.01	84.55	20.02	4.2
4 586.95	86.99	36 272.0	10 694.2	10.34	0.11	119.51	15.26	5.5
2 729.24	165.85	23 093.3	8 618.6	5.15	0.04	80.49	15.87	3.6
3 289.40	116.26	26 878.6	7 747.9	6.64	0.68	136.58	7.81	3.4
2 800.78	42.28	39 572.0	4 565.8	5.45	0.66	78.86	16.00	4.2
3 264.20	52.84	51 866.1	6 022.7	6.31	−0.10	136.58	17.44	3.6
3 453.62	165.04	58 749.8	3 721.1	6.35	−0.03	138.21	17.98	3.1
1 741.45	10.57	23 990.8	861.0	7.37	−1.63	75.61	20.99	1.6
2 035.75	13.82	25 694.9	3 571.5	8.39	−0.43	102.44	21.66	3.4
1 578.00	8.13	23 736.3	2 845.5	5.15	0.04	76.42	21.46	2.7
4 167.44	58.44	34 314.3	5 060.1	12.88	0.22	136.58	24.78	2.8
2 799.97	21.14	22 809.5	3 552.0	9.14	−0.74	88.62	24.96	3.9

se evitará incluir a las dos variables, Time y Accounts, en el modelo. También el coeficiente de correlación muestral entre Change y Rating, que es 0.549, es elevado y merece ser considerado más cuidadosamente.

Al observar los coeficientes de correlación muestrales entre Sales y cada una de las variables independientes se puede tener una rápida idea de cuáles de las variables independientes son, en sí mismas, buenos predictores. Se encuentra que el mejor predictor de Sales es Accounts, debi-

TABLA 16.6 DEFINICIÓN DE LAS VARIABLES EN LOS DATOS DE CRAVENS

Variable	Definición
Ventas	Total de ventas acreditadas al representante de ventas
Antigüedad (Time)	Antigüedad del empleado en meses
Potencial (Poten)	Potencial de mercado: ventas industriales totales en unidades en el territorio de ventas*
GastPubl (AdvExp)	Gastos del territorio en publicidad
Participación (Share)	Participación en el mercado: promedio ponderado de los últimos cuatro años
Cambio (Change)	Cambio, en los últimos cuatro años en participación en el mercado
Cuentas (Accounts)	Número de cuentas asignadas a los representantes de ventas*
Trabajo (Work)	Carga de trabajo: índice ponderado basado en compras anuales y concentración de cuentas
Evaluación (Rating)	Evaluación general del representante de ventas sobre ocho dimensiones de desempeño: una evaluación agregada en una escala del 1-7

*Estos datos fueron codificados para proteger la confidenciabilidad.

FIGURA 16.13 COEFICIENTES DE CORRELACIÓN MUESTRAL DE LOS DATOS DE CRAVENS

	Sales	Time	Poten	AdvExp	Share	Change	Accounts	Work
Time	0.623							
Poten	0.598	0.454						
AdvExp	0.596	0.249	0.174					
Share	0.484	0.106	-0.211	0.264				
Change	0.489	0.251	0.268	0.377	0.085			
Accounts	0.754	0.758	0.479	0.200	0.403	0.327		
Work	-0.117	-0.179	-0.259	-0.272	0.349	-0.288	-0.199	
Rating	0.402	0.101	0.359	0.411	-0.024	0.549	0.229	-0.277

do a que su coeficiente de correlación muestral es el más alto (0.754). Recuerde que si sólo se tiene una variable independiente, el cuadrado del coeficiente de correlación muestral es el coeficiente de determinación. Por tanto, Accounts explica $(0.754)^2(100)$, o 56.85%, de la variabilidad en Sales. Las variables independientes que siguen en importancia son Time, Poten y AdvExp, cada una con un coeficiente de correlación muestral de 0.6, aproximadamente.

Aun cuando existen problemas potenciales de multicolinealidad, se va a obtener una ecuación estimada de regresión con estas ocho variables. Con el paquete de software Minitab se obtienen los resultados que se presentan en la figura 16.14. El coeficiente de determinación ajustado para este modelo de regresión múltiple con ocho variables es 88.3%. Observe, sin embargo, que los valores-*p* de las pruebas *t* para cada uno de los parámetros indican que sólo Poten, AdvExp y Share son significativos a un nivel de significancia $\alpha = 0.05$, dado el efecto de todas las demás variables. Por tanto, se deseará investigar los resultados que se obtienen si se usan solamente estas tres

FIGURA 16.14 RESULTADOS DE MINITAB PARA EL MODELO CON OCHO VARIABLES INDEPENDIENTES

```
The regression equation is
Sales = - 1508 + 2.01 Time + 0.0372 Poten + 0.151 AdvExp + 199 Share
        + 291 Change + 5.55 Accounts + 19.8 Work + 8 Rating

Predictor       Coef    SE Coef       T      p
Constant      1507.8      778.6   -1.94  0.071
Time           2.010      1.931    1.04  0.313
Poten       0.037205   0.008202    4.54  0.000
AdvExp       0.15099    0.04711    3.21  0.006
Share         199.02      67.03    2.97  0.009
Change         290.9      186.8    1.56  0.139
Accounts       5.551      4.776    1.16  0.262
Work           19.79      33.68    0.59  0.565
Rating           8.2      128.5    0.06  0.950

S = 449.0   R-sq = 92.2%   R-sq(adj) = 88.3%

Analysis of Variance

SOURCE           DF         SS        MS       F      p
Regression        8   38153568   4769196   23.65  0.000
Residual Error   16    3225984    201624
Total            24   41379552
```

FIGURA 16.15 RESULTADOS DE MINITAB PARA EL MODELO CON TRES VARIABLES, POTEN, ADVEXP Y Share

```
The regression equation is
Sales = - 1604 + 0.0543 Poten + 0.167 AdvExp + 283 Share

Predictor       Coef    SE Coef       T       p
Constant     -1603.6      505.6   -3.17   0.005
Poten       0.054286   0.007474    7.26   0.000
AdvExp       0.16748    0.04427    3.78   0.001
Share         282.75      48.76    5.80   0.000

S = 545.5   R-sq = 84.9%   R-sq(adj) = 82.7%

Analysis of Variance

SOURCE           DF         SS         MS       F       p
Regression        3   35130240   11710080   39.35   0.000
Residual Error   21    6249310     297586
Total            24   41379552
```

variables. En la figura 16.15 se muestran los resultados proporcionados por Minitab para la ecuación estimada de regresión con estas tres variables. Se ve que el coeficiente de determinación ajustado para esta ecuación estimada de regresión es 82.7%, el cual, aunque no es tan bueno como el de la ecuación estimada de regresión con ocho variables, es alto.

¿Cómo se puede encontrar una ecuación estimada de regresión que dé mejores resultados, dada la información de que se dispone? Una posibilidad es calcular todas las regresiones posibles. Es decir, obtener ocho ecuaciones estimadas de regresión con una sola variable (cada una de las cuales corresponde a una de las variables independientes), 28 ecuaciones estimadas de regresión con dos variables independientes (que es el número de combinaciones de ocho variables tomadas de dos en dos), y así sucesivamente. Para los datos de Cravens se necesitan, en total, 255 ecuaciones estimadas de regresión conteniendo una o más de las variables independientes.

Con los excelentes paquetes de software de que se dispone en la actualidad, se pueden calcular todas estas regresiones. Sin embargo, hacerlo representa una gran cantidad de cálculos y requiere que se revise una gran cantidad de resultados de computadora, la mayor parte de los cuales corresponderán a modelos obviamente pobres. En lugar de hacer esto se prefiere seguir un método más sistemático para elegir el subconjunto de variables independientes que proporcione la mejor ecuación estimada de regresión. En la sección siguiente se presentan algunos de los métodos más conocidos.



Procedimientos de elección de variables

Los procedimientos de selección de variables son especialmente útiles en las primeras etapas de la construcción de un modelo, pero no pueden sustituir la experiencia y el criterio del analista.

En esta sección se verán cuatro **procedimientos de selección de variables**: la regresión por pasos, la selección hacia adelante, la selección hacia atrás y la regresión del mejor subconjunto. Dado un conjunto de datos en el que hay varias variables independientes, estos procedimientos permiten determinar con qué variables independientes se obtiene el mejor modelo. Los tres primeros procedimientos son iterativos; en cada paso del procedimiento se agrega o se elimina una variable independiente y se evalúa el nuevo modelo. El procedimiento continúa hasta que un criterio de detención indica que el procedimiento ya no puede hallar un modelo mejor. El último procedimiento (mejores subconjuntos) no es un procedimiento que evalúe las variables de una en una, sino que evalúa modelos de regresión en los que intervienen distintos subconjuntos de variables independientes.

En los procedimientos regresión por pasos, selección hacia adelante y eliminación hacia atrás, en cada paso, el criterio para elegir una variable independiente para agregarla o eliminarla del modelo, se basa en el estadístico F presentado en la sección 16.2. Suponga, por ejemplo, que se desea considerar si agregar x_2 a un modelo en el que interviene x_1 o eliminar x_2 de un modelo en el que intervienen x_1 y x_2. Para probar si la adición o la eliminación de x_2 es estadísticamente significativa, las hipótesis nula y alternativa pueden establecerse como sigue:

$$H_0\colon \beta_2 = 0$$
$$H_a\colon \beta_2 \neq 0$$

En la sección 16.2 (ver ecuación (16.10)) se mostró que

$$F = \frac{\dfrac{\text{SCE}(x_1) - \text{SCE}(x_1, x_2)}{1}}{\dfrac{\text{SCE}(x_1, x_2)}{n - p - 1}}$$

es usado como criterio para determinar si la presencia de x_2 en el modelo causa una reducción significativa de la suma de los cuadrados debidos al error. El valor-p correspondiente a este estadístico F es el criterio que se emplea para determinar si se debe agregar o eliminar una variable independiente del modelo de regresión. Para el rechazo se emplea la regla usual: rechazar H_0 si valor-$p \leq \alpha$.

Regresión por pasos

El procedimiento de regresión por pasos empieza por determinar en cada paso si alguna de las variables *que ya se encuentran en el modelo* debe ser eliminada. Para esto primero se calcula el estadístico F y el correspondiente valor-p para cada una de las variables independientes que intervienen en el modelo. Minitab le llama al nivel de significancia α que se emplea para determinar si una variable independiente debe ser eliminada del modelo *Alpha to remove* (*Alpha para eliminar*). Si el valor-p de alguna de las variables independientes es mayor que *Alpha to remove,* la variable independiente que tenga el mayor valor-p se elimina del modelo y el proceso de regresión por pasos empieza un nuevo paso.

Si ninguna de las variables independientes puede ser eliminada del modelo, el procedimiento trata de ingresar otra variable independiente al modelo. Para hacer esto primero se calcula el estadístico F y el valor-p de cada variable independiente que no está en el modelo. Minitab le llama al nivel de significancia α que emplea para determinar si una variable independiente debe agregarse al modelo, *Alpha to enter* (*Alpha para ingresar*). La variable independiente que tiene el menor valor-p es ingresada al modelo siempre que su valor-p sea menor que *Alpha to enter.* Este procedimiento continúa de la misma forma hasta que no haya ninguna variable independiente que pueda ser eliminada o agregada al modelo.

En la figura 16.16 se muestran los resultados obtenidos por Minitab con el procedimiento de regresión por pasos aplicado a los datos de Cravens, con 0.05 como *Alpha to remove* y 0.05 como *Alpha to enter.* Este procedimiento por pasos terminó en cuatro pasos. La ecuación estimada de regresión obtenida con el procedimiento de regresión por pasos de Minitab es

$$\hat{y} = -1\,441.93 + 9.2 \text{ Accounts (Cuentas)} +$$
$$0.175 \text{ AdvExp (GastPubl)} + 0.0382 \text{ Poten} + 190 \text{ Share (Participación)}$$

Dado que los procedimientos de una en una variable no consideran todos los subconjuntos posibles de una cantidad dada de variables independientes, estos procedimientos no necesariamente eligen el modelo con el que se obtenga el valor mayor R-sq.

En la figura 16.16, observe también que, después de cuatro pasos, $s = \sqrt{\text{CME}}$ se ha reducido de 881 en el mejor modelo con una variable [Cuentas (Accounts)] a 454. El valor de R-sq ha aumentado de 56.85 a 90.04% y el R-sq(adj) de la ecuación estimada de regresión recomendada es 88.05%.

En resumen, en cada paso del procedimiento de regresión por pasos, lo primero que se considera es si alguna de las variables independientes puede ser eliminada del modelo que se tiene. Si

FIGURA 16.16 RESULTADO DE LA REGRESIÓN POR PASOS DE MINITAB PARA LOS DATOS DE CRAVENS

```
Alpha-to-Enter: 0.05     Alpha-to-Remove: 0.05

Response is Sales on 8 predictors, with N = 25

    Step          1        2          3           4
Constant     709.32    50.29    -327.24    -1441.93

Accounts       21.7     19.0       15.6         9.2
T-Value        5.50     6.41       5.19        3.22
P-Value       0.000    0.000      0.000       0.004

AdvExp                 0.227      0.216       0.175
T-Value                 4.50       4.77        4.74
P-Value                0.000      0.000       0.000

Poten                            0.0219      0.0382
T-Value                            2.53        4.79
P-Value                           0.019       0.000

Share                                           190
T-Value                                        3.82
P-Value                                       0.001

S               881      650        583         454
R-Sq          56.85    77.51      82.77       90.04
R-Sq(adj)     54.97    75.47      80.31       88.05
C-p            67.6     27.2       18.4         5.4
```

ninguna de las variables independientes puede ser eliminada del modelo, el procedimiento verifica si alguna de las variables independientes que no intervienen en el modelo puede ser ingresada al modelo. Debido a la naturaleza del procedimiento de regresión por pasos, puede ser que una variable independiente sea ingresada al modelo en un paso, en un paso subsiguiente eliminada y después ingresada al modelo en un paso posterior. El procedimiento se detiene cuando no hay ya ninguna variable independiente que pueda ser eliminada del modelo ni agregada al modelo.

Selección hacia adelante

En el procedimiento de selección hacia adelante se empieza sin ninguna variable independiente y se van agregando variables de una en una con el mismo procedimiento que se usa en la regresión por pasos para determinar si una variable independiente debe ser ingresada al modelo. Pero, en el procedimiento de selección hacia adelante no se permite que se elimine del modelo una variable que ha sido ingresada. El procedimiento se detiene cuando el valor-p de cada una de las variables independientes que no están en el modelo es mayor que *Alpha to enter.*

La ecuación estimada de regresión obtenida mediante el procedimiento de selección hacia adelante de Minitab es

$$\hat{y} = -1\,441.93 + 9.2 \text{ Accounts (Cuentas)} + 0.175$$
$$\text{AdvExp (GastPubl)} + 0.0382 \text{ Poten} + 190 \text{ Share (Participación)}$$

Por tanto, en el caso de los datos de Cravens, con el procedimiento de selección hacia adelante (con 0.05 como *Alpha to enter*) se llega a la misma ecuación estimada de regresión que con el procedimiento por pasos.

Eliminación hacia atrás

En el procedimiento de eliminación hacia atrás se empieza con un modelo en el que se incluyen todas las variables independientes. Después, de una en una, se van eliminando variables independientes mediante el mismo procedimiento que en la regresión por pasos. Sin embargo, en el procedimiento de eliminación hacia atrás no se permite que una variable que ya ha sido eliminada vuelva a ser ingresada al modelo. El procedimiento se detiene cuando ninguna de las variables independientes del modelo tenga un valor-p mayor que *Alpha to remove*.

La ecuación estimada de regresión obtenida con el procedimiento de eliminación hacia atrás de Minitab aplicado a los datos de Cravens (con 0.05 como *Alpha to remove*) es

$$\hat{y} = -1\,312 + 3.8 \text{ Time (Antigüedad)} + 0.0444$$
$$\text{Potencial} + 0.152 \text{ AdvExp (GastPubl)} + 259 \text{ Share (participación)}$$

Al comparar la ecuación estimada de regresión obtenida mediante el procedimiento de eliminación hacia atrás con la ecuación estimada de regresión obtenida con el procedimiento de selección hacia adelante, se ve que hay tres variables independientes comunes a los dos procedimientos: AdvExp, Poten y Share. Pero, en el procedimiento de eliminación hacia atrás se incluyó Time en lugar de Accounts.

La selección hacia adelante y la eliminación hacia atrás pueden llevar a modelos distintos.

La selección hacia adelante y la eliminación hacia atrás son dos extremos en la construcción de modelos; en el procedimiento de selección hacia adelante se empieza sin ninguna variable independiente en el modelo y, una por una, se van agregando variables independientes, mientras que en el procedimiento de eliminación hacia atrás se empieza teniendo todas las variables independientes en el modelo y, de una en una, se eliminan variables. Con los dos procedimientos se puede llegar a la misma ecuación estimada de regresión. Sin embargo, también es posible que se llegue a ecuaciones estimadas de regresión diferentes, como ocurre en el caso de los datos de Cravens. ¿Por cuál de las ecuaciones estimadas de regresión decidirse? Esto es algo que queda a discusión. Al final el analista tiene que aplicar su propio criterio. El procedimiento de los mejores subconjuntos para la construcción de modelos que se estudia a continuación proporciona más información para la construcción de modelos, información que debe ser considerada antes de tomar la decisión final.

Regresión de los mejores subconjuntos

La regresión por pasos, la selección hacia adelante y la eliminación hacia atrás son métodos para elegir un modelo de regresión que agrega o elimina, una por una, variables independientes. Ninguno de estos métodos garantiza que dado un determinado número de variables se encuentre el mejor modelo. Por tanto, estos métodos de una por una suelen ser considerados como heurísticos para la selección de un buen modelo de regresión.

Algunos paquetes de software usan un procedimiento llamado regresión de los mejores subconjuntos que permite al usuario, hallar el mejor modelo de regresión para un número determinado de variables independientes. Minitab cuenta con este procedimiento. En la figura 16.17 se presenta parte de los resultados obtenidos mediante el procedimiento de los mejores subconjuntos de Minitab aplicado a los datos de Cravens.

En estos resultados aparecen las dos mejores ecuaciones de regresión estimada con una sola variable, las dos mejores ecuaciones con dos variables, las dos mejores ecuaciones con tres variables, etc. El criterio que se emplea para determinar cuáles son las mejores ecuaciones estimadas de regresión con un determinado número de predictores es el valor del coeficiente de determinación (R-sq). Por ejemplo, Accounts proporciona la mejor ecuación estimada de regresión con una sola variable independiente, R-sq = 56.8%; al usar AdvExp y Accounts se obtiene la mejor ecuación estimada de regresión con dos variables independientes, R-sq = 77.5%, y con Poten, AdvExp y Share se obtiene la mejor ecuación estimada de regresión con tres variables independientes, R-sq = 84.9%. Para los datos de Cravens, el mayor coeficiente de determinación ajustado (Adj.R-sq = 89.4%) es el del modelo con seis variables independientes, Time, Poten, AdvExp, Share, Change y Accounts. Sin embargo, el coeficiente de determinación ajustado del mejor modelo con cuatro variables independientes (Poten, AdvExp, Share y Accounts) es casi igual de alto (88.1%). Por lo general, se prefiere el modelo más sencillo con el menor número de variables.

FIGURA 16.17 PARTE DE LOS RESULTADOS OBTENIDOS CON LA REGRESIÓN DE LOS MEJORES SUBCONJUNTOS DE MINITAB

```
                                            A
                                            c
                                      A   C c   R
                                    P d S h o   a
                                  T o v h a u W t
                                  i t E a n n o I
                Adj.              m e x r g t r n
Vars    R-sq    R-sq         s    e n p e e s K g

   1    56.8    55.0    881.09              X
   1    38.8    36.1    1049.3    X
   2    77.5    75.5    650.39        X     X
   2    74.6    72.3    691.11      X     X
   3    84.9    82.7    545.52      X X X
   3    82.8    80.3    582.64      X X     X
   4    90.0    88.1    453.84      X X X   X
   4    89.6    87.5    463.93    X X X X
   5    91.5    89.3    430.21    X X X X X
   5    91.2    88.9    436.75      X X X X X
   6    92.0    89.4    427.99    X X X X X X
   6    91.6    88.9    438.20      X X X X X X
   7    92.2    89.0    435.66    X X X X X X X
   7    92.0    88.8    440.29    X X X X X X   X
   8    92.2    88.3    449.02    X X X X X X X X
```

Elección final

El análisis de los datos de Cravens hecho hasta ahora es una buena preparación para tomar una decisión por un modelo, pero antes habrá que hacer también otros análisis. Como se indicó en los capítulos 14 y 15, es necesario hacer un cuidadoso análisis de los residuales. Se desea que la gráfica de los residuales del modelo elegido parezca una banda horizontal. Suponga que en los residuales no se encuentre ningún problema y que se desee emplear los resultados del procedimiento de los mejores subconjuntos para decidirse por un modelo.

El procedimiento de los mejores subconjuntos indica que el mejor modelo con cuatro variables es el que contiene las variables independientes Poten, AdvExp, Share y Accounts. Este modelo resulta ser también el modelo con cuatro variables encontrado mediante el procedimiento de regresión por pasos. La tabla 16.7 ayuda a tomar la decisión final. En esta tabla se muestran varios modelos que contienen algunas, o las cuatro, de estas cuatro variables independientes.

TABLA 16.7 MODELOS SELECCIONADOS CON Accounts, Poten, AdvExp Y Share

Modelo	Variables independientes	Adj. R-sq
1	Accounts	55.0
2	AdvExp, Accounts	75.5
3	Poten, Share	72.3
4	Poten, AdvExp, Accounts	80.3
5	Poten, AdvExp, Share	82.7
6	Poten, AdvExp, Share, Accounts	88.1

En la tabla 16.7 se ve que el modelo que sólo tiene AdvExp y Accounts es bueno. Su coeficiente de determinación ajustado es 75.5%, y con el modelo con las cuatro variables sólo se logra un aumento de 12.6 puntos porcentuales. El modelo más sencillo que sólo tiene dos variables puede preferirse si, por ejemplo, es difícil medir el potencial de mercado (Poten). Sin embargo, si ya se cuenta con los datos y se requiere gran precisión en la predicción de las ventas, es claro que se preferirá el modelo con las cuatro variables.

NOTAS Y COMENTARIOS

1. En el procedimiento por pasos se requiere que *Alpha to remove* sea mayor o igual que *Alpha to enter*. Este requerimiento evita que, en un mismo paso, una misma variable sea eliminada y reingresada.
2. Para crear nuevas variables independientes que puedan ser usadas con los procedimientos de esta sección se usan funciones de las variables independientes. Por ejemplo, si se desea tener en el modelo x_1x_2 para reflejar la interacción, se usan los datos de x_1 y de x_2 para crear una variable $z = x_1x_2$.
3. Ninguno de los procedimientos que agregan o eliminan variables de una en una garantizan que se encuentre el mejor modelo de regresión. Pero estos procedimientos son excelentes para hallar buenos modelos, en especial cuando hay poca multicolinealidad.

Ejercicios

Aplicaciones

16. En un estudio se obtuvieron datos de variables que pueden estar relacionadas con el número de semanas que está desempleado un trabajador de la industria. La variable dependiente de este estudio (semanas) se definió como el número de semanas que un empleado está desempleado debido a despido. En este estudio se usaron las siguientes variables independientes.



Age (edad)	Edad del trabajador
Educ (educación)	Número de años de estudio
Married (casado)	Variable ficticia; 1 si está casado, 0 si no es así
Head (cabeza)	Variable ficticia; 1 si es cabeza de familia, 0 si no es así
Tenure (ocupación)	Número de años en el trabajo anterior
Manager (administrativo)	Variable ficticia; 1 si su ocupación es en administración , 0 si no es así
Sales (ventas)	Variable ficticia; 1 si su ocupación es en ventas, 0 si no es así

Estos datos se encuentran en el archivo Layoffs del disco compacto que se distribuye con el libro.

a. Obtenga la mejor ecuación estimada de regresión que tenga una variable.
b. Emplee el procedimiento por pasos para obtener la mejor ecuación estimada de regresión. Use 0.05 como *Alpha to enter* y *Alpha to remove*.
c. Use el procedimiento de selección hacia adelante para obtener la mejor ecuación estimada de regresión. Use 0.05 como *Alpha to enter.*
d. Use el procedimiento de eliminación hacia atrás para obtener la mejor ecuación estimada de regresión. Use 0.05 como *Alpha to remove.*
e. Use el procedimiento de regresión de los mejores subconjuntos para obtener la mejor ecuación estimada de regresión.



17. La Ladies Professional Golfers Association (LPGA) lleva estadísticas sobre el desempeño y las ganancias de sus miembros en la LPGA Tour. En el archivo titulado LPGATour 2 del disco compacto se presentan las estadísticas de fin de año sobre el desempeño de las 30 jugadoras que tuvieron las mejores ganancias en la LPGA Tour de 2005 (www.lpga.com, 2006). Earnings (ga-

nancias) (miles) son las ganancias totales en miles de dólares en todos los eventos de la gira; Scoring Avg., es la puntuación promedio de la jugadora en todos los eventos; Drive Average es la distancia media alcanzada en el drive por el jugador en yardas; Greens in Reg., es el porcentaje de veces que una jugadora llega al green en regulación; Putting Avg., es el promedio de putts realizados en el green en regulación, y Sand Saves es el porcentaje de veces que la jugadora logra "subir y bajar" ("up and down") cuando se encuentra en un búnker de arena al lado del green. Sea Drive Greens una nueva variable independiente que represente la interacción entre la distancia media alcanzada en el drive por el jugador y Greens in Reg. Use los métodos de esta sección para obtener la mejor ecuación estimada de regresión múltiple para estimar Scoring Avg de un jugador.

18. Jeff Sagarin proporciona, desde 1985, evaluaciones deportivas para *USA Today*. En el béisbol sus pronósticos RPG (runs/game) estadísticos toman en cuenta todas las estadísticas de ofensiva del jugador y, se asegura, que es la mejor medida del verdadero valor de la ofensiva de un jugador. En los datos que se presentan a continuación se da el RPG y varios estadísticos de ofensiva de la temporada de la Liga Mayor de Béisbol correspondientes a 20 miembros de los Yankees de Nueva York (www.usatoday.com, 3 de marzo de 2006). Los rótulos de las columnas se definen como sigue: RPG, estadístico que predice número de carreras por juego; H, batazos buenos; 2B, dobles; 3B, triples; HR, cuadrangulares; RBI, carreras bateadas; BB, bases por bola; SO, ponchadas; SB, bases robadas; CS, atrapado en robo de base; OBP, porcentaje en base; SLG, porcentaje de potencia de bateo; AVG, promedio de bateo.



Jugador	RPG	H	2B	3B	HR	RBI	BB	SO	SB	CS	OBP	SLG	AVG
D Jeter	6.51	202	25	5	19	70	77	117	14	5	0.389	0.45	0.309
H Matsui	6.32	192	45	3	23	116	63	78	2	2	0.367	0.496	0.305
A Rodriguez	9.06	194	29	1	48	130	91	139	21	6	0.421	0.61	0.321
G Sheffield	6.93	170	27	0	34	123	78	76	10	2	0.379	0.512	0.291
R Cano	5.01	155	34	4	14	62	16	68	1	3	0.32	0.458	0.297
B Williams	4.14	121	19	1	12	64	53	75	1	2	0.321	0.367	0.249
J Posada	5.36	124	23	0	19	71	66	94	1	0	0.352	0.43	0.262
J Giambi	9.11	113	14	0	32	87	108	109	0	0	0.44	0.535	0.271
T Womack	2.91	82	8	1	0	15	12	49	27	5	0.276	0.28	0.249
T Martinez	5.08	73	9	0	17	49	38	54	2	0	0.328	0.439	0.241
M Bellhorn	4.07	63	20	0	8	30	52	112	3	0	0.324	0.357	0.21
R Sierra	3.27	39	12	0	4	29	9	41	0	0	0.265	0.371	0.229
J Flaherty	1.83	21	5	0	2	11	6	26	0	0	0.206	0.252	0.165
B Crosby	3.48	27	0	1	1	6	4	14	4	1	0.304	0.327	0.276
M Lawton	5.15	6	0	0	2	4	7	8	1	0	0.263	0.25	0.125
R Sanchez	3.36	12	1	0	0	2	2	3	0	1	0.326	0.302	0.279
A Phillips	2.13	6	4	0	1	4	1	13	0	0	0.171	0.325	0.15
M Cabrera	1.19	4	0	0	0	0	0	2	0	0	0.211	0.211	0.211
R Johnson	3.44	4	2	0	0	0	1	4	0	0	0.3	0.333	0.222
F Escalona	5.31	4	1	0	0	2	1	4	0	0	0.375	0.357	0.286

Considere que la variable dependiente es la estadística RPG.

a. Obtenga la mejor ecuación estimada de regresión con una variable.
b. Emplee los métodos de esta sección para obtener la mejor ecuación estimada de regresión múltiple que estime el RPG de un jugador.



19. Vaya al ejercicio 14. Mediante la edad, la presión sanguínea, si la persona es o no fumadora y cualquier interacción entre estas variables, obtenga una ecuación estimada de regresión que sirva para predecir riesgo. Haga una descripción breve del proceso que utilice para obtener esta ecuación estimada de regresión para estos datos.



Método de regresión múltiple para el diseño de experimentos

En la sección 15.7 se vio el uso de las variables ficticias en el análisis de regresión múltiple. En esta sección se muestra cómo el uso de variables ficticias en una ecuación de regresión múltiple puede proporcionar otro método para resolver problemas de diseño experimental (o diseño de experimentos). El uso de la regresión múltiple en el diseño experimental se demostrará con el ejemplo del diseño completamente aleatorizado presentado en el capítulo 13.

Recuerde que Chemitech elaboró un nuevo sistema de filtración para el suministro público de agua. Chemitech compraría los componentes del sistema de filtración a diversos proveedores y los armaría en sus instalaciones en Columbia, Carolina del Sur. Se tenían tres métodos de ensamblados, identificados como método A, B y C. Los gerentes de Chemitech deseaban saber qué método de ensamblado producía mayor número de sistemas de filtración por semana.

Se tomó una muestra aleatoria de 15 empleados y cada uno de los tres métodos de ensamblado le fue asignado aleatoriamente a 5 de estos empleados. En la tabla 16.8 se presenta el número de unidades ensambladas por cada empleado. Las medias muestrales del número de unidades producidas con cada uno de los tres métodos son las siguientes:

Método de ensamblado	Número medio producido
A	62
B	66
C	52

Aunque el método B parece ser el que proporciona una tasa de producción más alta, lo que interesa saber es si las tres medias muestrales observadas son suficientemente diferentes como para poder concluir que las medias poblacionales correspondientes a los tres métodos de ensamblado son diferentes.

En el método de regresión aplicado a este problema se empieza por definir las variables ficticias que se usarán para indicar cuál de los métodos de ensamblado fue usado. Como en el problema de Chemitech hay tres métodos de ensamblado, o tratamientos, se necesitan dos variables ficticias. En general, si el factor que se va a investigar tiene k niveles, o tratamientos, se necesita definir $k - 1$ variables ficticias. Para el experimento de Chemitech se definen las variables ficticias A y B de la manera que se muestra en la tabla 16.9.

TABLA 16.8 NÚMERO DE UNIDADES PRODUCIDAS POR LOS 15 TRABAJADORES

Método		
A	**B**	**C**
58	58	48
64	69	57
55	71	59
66	64	47
67	68	49

TABLA 16.9 VARIABLES FICTICIAS PARA EL EXPERIMENTO DE CHEMITECH

A	B	
1	0	Observación relacionada con el método de ensamblado A
0	1	Observación relacionada con el método de ensamblado B
0	0	Observación relacionada con el método de ensamblado C

Las variables ficticias se pueden usar para relacionar el número de unidades, y, producidas por semana con el método de ensamblado usado por el empleado.

$$
\begin{aligned}
E(y) &= \text{Valor esperado del número de unidades producidas por semana} \\
&= \beta_0 + \beta_1\text{A} + \beta_2\text{B}
\end{aligned}
$$

Por tanto, si interesa el valor esperado del número de unidades ensambladas por semana por un empleado mediante el método C, de acuerdo con el procedimiento para asignar valores numéricos a las variables ficticias se tendrá A = B = 0. La ecuación de regresión múltiple se reduce entonces a

$$E(y) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

La interpretación es que β_0 se puede interpretar como el valor esperado de la cantidad de unidades ensambladas por semana por un empleado que use el método C. En otras palabras, β_0 es la media de la cantidad de unidades ensambladas por semana mediante el método C.

A continuación se considera la forma de la ecuación de regresión múltiple correspondiente a cada uno de los otros métodos. Los valores de las variables ficticias correspondientes al método A son A = 1 y B = 0, y entonces

$$E(y) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

Los correspondientes al método B son A = 0 y B = 1, y entonces

$$E(y) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

Como se ve, $\beta_0 + \beta_1$ representa la media del número de unidades ensambladas por semana mediante el método A, y $\beta_0 + \beta_2$ representa la media del número de unidades ensambladas por semana con el método B.

Ahora se desea obtener estimaciones para los coeficientes β_0, β_1 y β_2 y, de esta manera, obtener una estimación del número medio de unidades ensambladas por semana con cada uno de los métodos. En la tabla 16.10 se presentan los datos muestrales que consisten en 15 observaciones de A, B y y. En la figura 16.18 se presentan los resultados correspondientes obtenidos usando la regresión múltiple de Minitab. Como se ve, las estimaciones de β_0, β_1 y β_2 son $b_0 = 52$, $b_1 = 10$ y $b_2 = 14$. De esta manera, las mejores estimaciones de las medias del número de unidades ensambladas por semana con cada uno de los métodos de ensamblado son:

Método de ensamblado	**Estimación de $E(y)$**
A	$b_0 + b_1 = 52 + 10 = 62$
B	$b_0 = 52 + 14 = 66$
C	$b_0 = 52$

Observe que estas estimaciones, de los números medios de unidades producidas con cada uno de estos tres métodos de ensamblado, obtenidas mediante el análisis de regresión son las mismas que las medias muestrales presentadas previamente.

TABLA 16.10 DATOS PARA EL DISEÑO COMPLETAMENTE ALEATORIZADO DE CHEMITECH

A	B	y
1	0	58
1	0	64
1	0	55
1	0	66
1	0	67
0	1	58
0	1	69
0	1	71
0	1	64
0	1	68
0	0	48
0	0	57
0	0	59
0	0	47
0	0	49

Ahora se va a ver cómo usar los resultados del análisis de regresión múltiple para realizar la prueba ANOVA de la diferencia entre las medias de estos tres métodos. Primero, se observa que si las medias no difieren

$$E(y) \text{ para el método A} - E(y) \text{ para el método C} = 0$$
$$E(y) \text{ para el método B} - E(y) \text{ para el método C} = 0$$

Como β_0 es igual a $E(y)$ al emplear el método C y $\beta_0 + \beta_1$ es igual a $E(y)$ al emplear el método A, la primera diferencia es igual a $(\beta_0 + \beta_1) - \beta_0 = \beta_1$. Y como $\beta_0 + \beta_2$ es igual a $E(y)$ al emplear el método B, la segunda diferencia es igual a $(\beta_0 + \beta_2) - \beta_0 = \beta_2$. Se concluye que entre

FIGURA 16.18 RESULTADOS DE LA REGRESIÓN MÚLTIPLE PARA EL DISEÑO COMPLETAMENTE ALEATORIZADO DE CHEMITECH

```
The regression equation is
y = 52.0 + 10.0 A + 14.0 B

Predictor     Coef  SE Coef      T      P
Constant    52.000    2.380  21.84  0.000
A           10.000    3.367   2.97  0.012
B           14.000    3.367   4.16  0.001

S = 5.32291   R-Sq = 60.5%   R-Sq(adj) = 53.9%

Analysis of Variance

SOURCE          DF       SS      MS      F      P
Regression       2   520.00  260.00   9.18  0.004
Residual Error  12   340.00   28.33
Total           14   860.00
```

los tres métodos no hay diferencia si $\beta_1 = 0$ y $\beta_2 = 0$. Por tanto, la hipótesis nula en una prueba para la diferencia entre las medias se puede expresar como

$$H_0 : \beta_1 = \beta_2 = 0$$

Tome el nivel de significancia $\alpha = 0.05$. Recuerde que para probar este tipo de hipótesis nula acerca de la significancia de la relación de regresión se emplea la prueba F de significancia general. En el resultado de Minitab que se presenta en la figura 16.18 se observa que el valor-p correspondiente a $F = 9.18$ es 0.004. Como valor-$p = 0.004 < \alpha = 0.05$, $H_0 : \beta_1 = \beta_2 = 0$ se rechaza y se concluye que las medias de los tres métodos de ensamblado no son iguales. Como la prueba F indica que la relación de regresión múltiple es significativa, se puede realizar una prueba t para determinar la significancia de cada uno de los parámetros, β_1 y β_2. Con $\alpha = 0.05$, los valores-p, 0.012 y 0.001, que aparecen en los resultados de Minitab, indican que las hipótesis nulas $H_0 : \beta_1 = 0$ y $H_0 : \beta_2 = 0$, se pueden rechazar. Por tanto, ambos parámetros son estadísticamente significativos. Así, se concluye que las medias de los parámetros A y C son diferentes y que también las medias de los parámetros B y C son diferentes.

Ejercicios

Métodos



20. Considere un diseño completamente aleatorizado en el que haya cuatro tratamientos: A, B, C y D. Escriba la ecuación de regresión múltiple que sirva para analizar estos datos. Defina todas las variables.
21. Dé una ecuación de regresión múltiple que sirva para analizar los datos de un diseño de bloque aleatorizado que tenga tres tratamientos y dos bloques. Defina todas las variables.
22. Dé una ecuación de regresión múltiple que sirva para analizar los datos de un diseño bifactorial que tenga dos niveles para el factor A y tres niveles para el factor B. Defina todas las variables.

Aplicaciones



23. La empresa Jacobs Chemical desea estimar el tiempo promedio (en minutos) necesario para mezclar un lote de un material empleando máquinas provenientes de tres fabricantes diferentes. Para limitar los costos de la prueba se mezclaron cuatro lotes de material en las máquinas producidas por cada uno de los fabricantes. A continuación se presentan los tiempos requeridos.

Fabricante 1	Fabricante 2	Fabricante 3
20	28	20
26	26	19
24	31	23
22	27	22

a. Dé una ecuación de regresión múltiple que sirva para analizar estos datos.
b. Dé las mejores estimaciones de los coeficientes en su ecuación.
c. En términos de los coeficientes de las ecuaciones de regresión cuáles son las hipótesis a probar para ver si los tiempos son iguales con las máquinas de los tres fabricantes.
d. ¿Cuál es la conclusión que se obtiene con un nivel de significancia 0.05?

24. En la publicidad de cuatro pinturas diferentes se asegura que todas tienen el mismo tiempo de secado. Para comprobar esto se probaron cinco muestras de cada pintura.

Los tiempos de secado de cada muestra se presentan a continuación

Pintura 1	Pintura 2	Pintura 3	Pintura 4
128	144	133	150
137	133	143	142
135	142	137	135
124	146	136	140
141	130	131	153

a. Use $\alpha = 0.05$ para probar si existe una diferencia significativa entre los tiempos de secado.
b. Dé una estimación del tiempo medio de secado de la pintura 2. ¿Cómo se obtiene de los resultados de un paquete de software?

25. Un comerciante de automóviles realizó una prueba para determinar si el tiempo requerido para ajustar un motor dependía de si se empleaba un analizador computarizado o un analizador electrónico. Como el tiempo que se necesita para ajustar un motor depende de si se trata de un auto pequeño, mediano o grande, se usaron los tres tipos de automóviles como bloques del experimento. Los datos que se obtuvieron (en minutos) son los que se presentan a continuación.

		Automóvil		
		Pequeño	Mediano	Grande
Analizador	Computarizado	50	55	63
	Electrónico	42	44	46

Emplee $\alpha = 0.05$ para probar si hay diferencias significativas.

26. Una empresa de ventas por catálogo diseñó un experimento factorial para probar los efectos del tamaño de un anuncio publicitario y su diseño sobre el número (en miles) de catálogos solicitados. Se consideraron tres diseños y dos tamaños diferentes del anuncio publicitario. De éstos se obtuvieron los datos siguientes. Pruebe si hay efectos significativos debido al diseño, al tamaño o a interacciones. Use $\alpha = 0.05$.

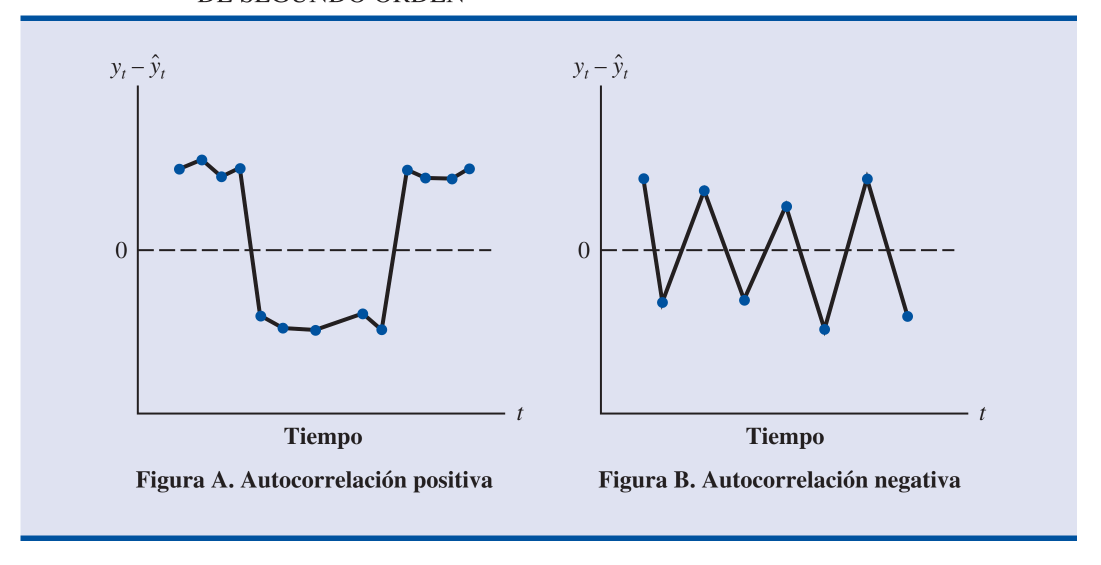
		Tamaño del anuncio publicitario	
		Pequeño	Grande
Diseño	A	8	12
		12	8
	B	22	26
		14	30
	C	10	18
		18	14



16.6 Autocorrelación y la prueba de Durbin-Watson

En los negocios y en la economía suele ocurrir que los datos que se usan en estudios de regresión estén correlacionados a lo largo del tiempo. No es raro que el valor de y en el periodo t, que se denota y_t esté relacionado con el valor de y en un periodo anterior. En tales casos se dice que existe **autocorrelación** en los datos (o **correlación serial**). Si el valor de y en el periodo t está rela-

FIGURA 16.19 DOS CONJUNTOS DE DATOS CON CORRELACIÓN DE SEGUNDO ORDEN



cionado con su valor en el periodo $t - 1$, existe correlación de primer orden. Si el valor de y en el periodo t está relacionado con su valor en el periodo $t - 2$, existe correlación de segundo orden y así sucesivamente.

Cuando hay autocorrelación se viola una de las suposiciones del modelo de regresión: los términos del error no son independientes. En el caso de la autocorrelación de primer orden, el error en el periodo t que se denota ϵ_t, estará relacionado con el error en el periodo $t - 1$, que se denota ϵ_{t-1}. En la figura 16.19 se ilustran dos casos de autocorrelación de primer orden. En la gráfica A se presenta el caso de una autocorrelación positiva; en la gráfica B el de una autocorrelación negativa. En la autocorrelación positiva se espera que el residual positivo de un periodo vaya seguido de un residual positivo en el periodo siguiente, que el residual negativo de un periodo vaya seguido de un residual negativo en el periodo siguiente y así sucesivamente. En la autocorrelación negativa se espera que el residual positivo de un periodo vaya seguido de un residual negativo en el periodo siguiente, después un residual positivo y así sucesivamente.

Si existe autocorrelación, se pueden cometer errores serios cuando se realizan pruebas de significancia estadística basadas en el modelo de regresión supuesto. Por tanto, es importante poder detectar la autocorrelación y tomar medidas correctivas. A continuación se mostrará cómo usar el estadístico de Durbin-Watson para detectar autocorrelación de primer orden.

Suponga que los valores de ϵ no sean independientes sino que estén relacionados de la manera siguiente:

$$\epsilon_t = \rho\epsilon_{t-1} + z_t \tag{16.16}$$

donde ρ es un parámetro cuyo valor absoluto es menor que 1 y z_i es una variable aleatoria distribuida normal e independientemente, que tienen media cero y varianza σ^2. En la ecuación 16.16 se ve que si $\rho = 0$, los términos del error no están relacionados y cada uno tiene media cero y varianza σ^2. En este caso no hay autocorrelación y se satisfacen las suposiciones de la regresión. Si $\rho > 0$, existe autocorrelación positiva; si $\rho < 0$, existe autocorrelación negativa. En cualquiera de estos casos, se violan las suposiciones de la regresión acerca del término del error.

En la prueba de **Durbin-Watson** para autocorrelación se usan los residuales para determinar si $\rho = 0$. Para simplificar la notación para el estadístico de Durbin-Watson el residual i se denota $e_i = y_i - \hat{y}_i$. El estadístico de prueba Durbin-Watson se calcula como sigue.

ESTADÍSTICO DE PRUEBA DURBIN-WATSON

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{16.17}$$

Si valores sucesivos de los residuales se encuentran cercanos unos de otros (autocorrelación positiva), el valor del estadístico de prueba Durbin-Watson será pequeño. Si valores sucesivos de los residuales se encuentran alejados unos de otros (autocorrelación negativa), el valor del estadístico de prueba Durbin-Watson será grande.

El estadístico de prueba Durbin-Watson va de cero a cuatro, si su valor es dos, esto indica que no existe autocorrelación. Durbin y Watson elaboraron tablas para determinar cuándo su estadístico indica la existencia de autocorrelación. En la tabla 16.11 se presentan límites inferiores y superiores (d_L y d_U) para las pruebas de hipótesis con $\alpha = 0.05$; n denota el número de observaciones. Siempre, la hipótesis nula a probar es que no existe autocorrelación

$$H_0: \rho = 0$$

La hipótesis alternativa que se prueba en la autocorrelación positiva es

$$H_a: \rho > 0$$

La hipótesis alternativa que se prueba en la autocorrelación negativa es

$$H_a: \rho < 0$$

TABLA 16.11 VALORES CRÍTICOS PARA LA PRUEBA DE DURBIN-WATSON PARA AUTOCORRELACIÓN

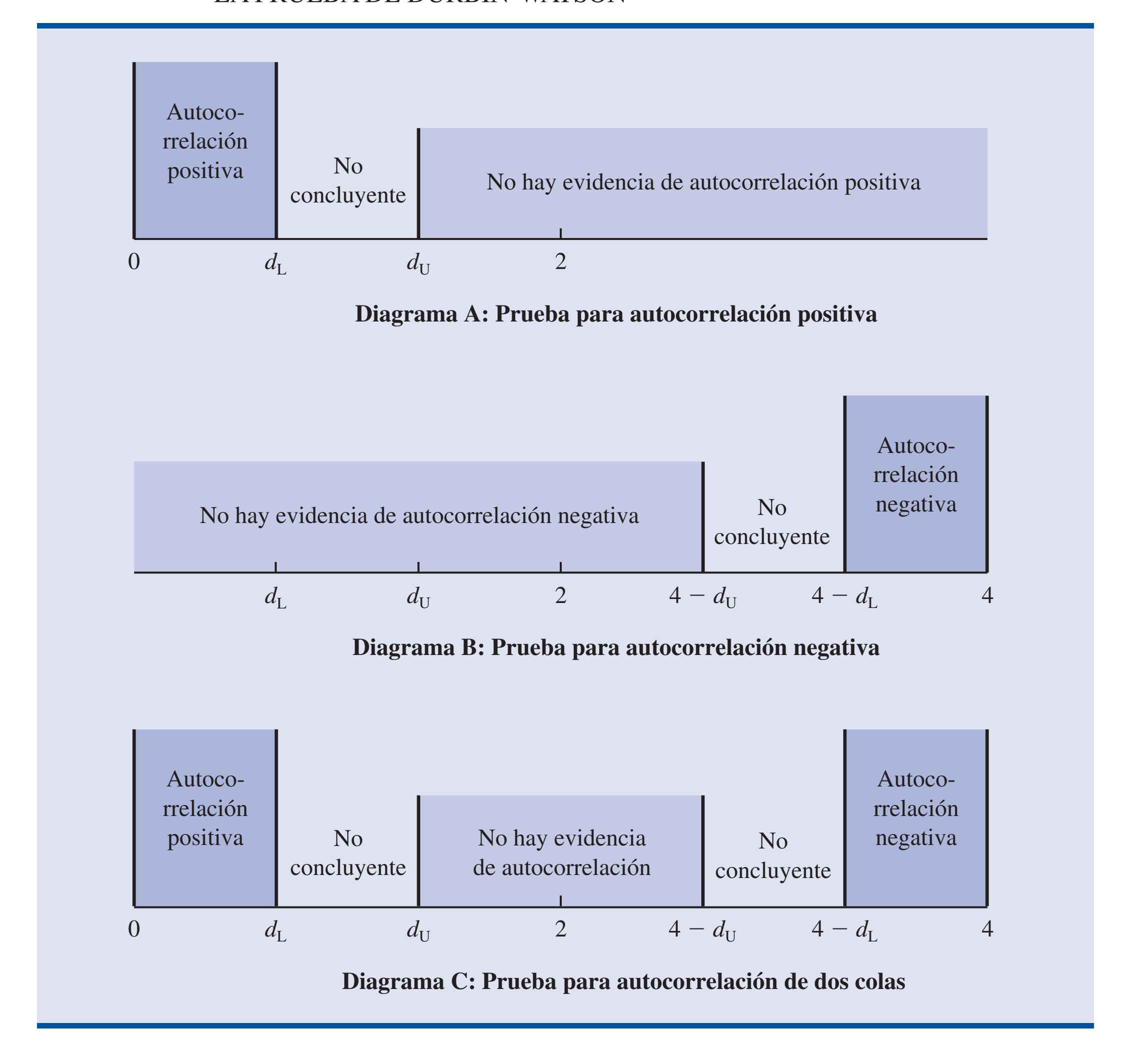
Nota: Los valores que se presentan en la tabla son los valores críticos para la prueba de Durbin-Watson de una cola para autocorrelación. En pruebas de dos colas, se duplica el nivel de significancia.

Puntos de significancia de d_L y d_U: $\alpha = 0.05$
Número de variables independientes

	1		2		3		4		5	
n*	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

* Para valores intermedios de *n*, interpolar linealmente.

FIGURA 16.20 PRUEBA DE HIPÓTESIS PARA AUTOCORRELACIÓN MEDIANTE LA PRUEBA DE DURBIN-WATSON



También se puede hacer una prueba de dos colas. En este caso la hipótesis alternativa es

$$H_a: \rho \neq 0$$

En la figura 16.20 se muestra el uso de los valores d_L y d_U de la tabla 16.11 para probar si existe autocorrelación. En el diagrama A se ilustra la prueba para autocorrelación positiva: Si $d < d_L$, se concluye que existe correlación positiva. Si $d_L \leq d \leq d_U$, se dice que la prueba no es concluyente. Si $d > d_U$, se concluye que no hay evidencia de autocorrelación positiva.

En el diagrama B se ilustra la prueba para autocorrelación negativa. Si $d > 4 - d_L$, se concluye que existe autocorrelación negativa. Si $4 - d_U \leq d \leq 4 - d_L$, se dice que la prueba no es concluyente. Si $d < 4 - d_U$, se concluye que no existe evidencia de autocorrelación negativa.

El diagrama C ilustra la prueba de dos colas. Si $d < d_L$ o $d > 4 - d_L$, se rechaza H_0 y se concluye que existe correlación. Si $d_L \leq d \leq d_U$ o si $4 - d_U \leq d \leq 4 - d_L$, se dice que la prueba no es concluyente. Si $d_U < d < 4 - d_U$, se concluye que no hay evidencia de autocorrelación.

Si se determina que hay una autocorrelación significativa, se debe verificar si se omitieron una o varias variables independientes importantes que tengan un efecto de orden temporal sobre la variable dependiente. Si no se encuentran tales variables, incluir una variable independiente que mida el tiempo en el que se hace la observación (el valor de esta variable, puede ser, por ejemplo, 1 para la primera observación, 2 para la segunda, etc.) algunas veces ayuda para eliminar o reducir la autocorrelación. Cuando no funcionan estos intentos para eliminar o reducir la autocorrelación, hacer transformaciones de las variables independientes resulta útil; un estudio sobre esas transformaciones puede encontrarse en libros más avanzados sobre análisis de regresión.

Observe que en las tablas de Durbin-Watson el menor valor para el tamaño de la muestra es 15. La razón es que para muestras menores, la prueba suele ser no concluyente; en realidad, se suele creer que el tamaño de la muestra debe ser de por lo menos 50 para que con la prueba se obtengan resultados que valgan la pena.

Ejercicios

Aplicaciones

27. En los datos siguientes se muestran los precios diarios de cierre (en dólares por acción) de IBM desde el 3 de noviembre de 2005, hasta el 1 de diciembre de 2005 (*Compustat,* 26 de febrero de 2006).



IBM

Fecha	Precio ($)
Nov. 3	82.87
Nov. 4	83.00
Nov. 7	83.61
Nov. 8	83.15
Nov. 9	82.84
Nov. 10	83.99
Nov. 11	84.55
Nov. 14	84.36
Nov. 15	85.53
Nov. 16	86.54
Nov. 17	86.89
Nov. 18	87.77
Nov. 21	87.29
Nov. 22	87.99
Nov. 23	88.80
Nov. 25	88.80
Nov. 28	89.11
Nov. 29	89.10
Nov. 30	88.90
Dic. 1	89.21

a. Defina la variable independiente Periodo, donde Periodo = 1 corresponda al dato del 3 de noviembre, Periodo = 2 corresponda al dato del 4 de Noviembre, etc. Obtenga una ecuación estimada de regresión que sirva para predecir el precio del cierre dado el valor del Periodo.
b. Emplee como nivel de significancia 0.05 y pruebe si existe autocorrelación positiva en estos datos.

28. Remítase al conjunto de datos de Craven de la tabla 16.5. En la sección 16.3 se mostró que el coeficiente de determinación ajustado de la ecuación estimada de regresión que contenía Accounts-(Cuentas), AdvExp (GastPubl), Poten y Share (Participación) era 88.1%. Use 0.05 como nivel de significancia y aplique la prueba de Durbin-Watson para determinar si existe autocorrelación positiva.

Resumen

En este capítulo se analizaron varios de los conceptos que se usan en la construcción de modelos para hallar la ecuación estimada de regresión. Primero se presentó el concepto de modelo lineal general para mostrar cómo pueden extenderse los métodos estudiados en los capítulos 14 y 15 a las relaciones curvilíneas y a los efectos de interacción. Después, se vio cómo emplear transformaciones a la variable dependiente cuando se presentan problemas como el de una varianza no constante en los términos del error.

En muchas aplicaciones del análisis de regresión se emplea un gran número de variables independientes. Para agregar o eliminar variables a un modelo de regresión se vio un método general basado en el estadístico F. Después se presentó un problema más grande en el que se tenían 25 observaciones y ocho variables independientes. También se vio que cuando se tienen problemas más grandes, uno de los asuntos a resolver es hallar el mejor subconjunto de variables independientes; para esto existen varios procedimientos de selección de variables: regresión por pasos, selección hacia adelante, eliminación hacia atrás y regresión de los mejores subconjuntos.

En la sección 16.5 se amplió el estudio para ver cómo obtener modelos de regresión múltiple que sirven como otro método para la solución de problemas de análisis de varianza y de diseño de experimentos. El capítulo concluyó con una aplicación del análisis de residuales mediante la prueba de Durbin-Watson para autocorrelación.

Glosario

Modelo lineal general Modelo de la forma $y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \epsilon$, en donde cada una de las variables independientes z_j $(j = 1, 2, \ldots, p)$ es función de $x_1, x_2, \ldots, x_k$, las variables para las que se han recolectado datos.

Interacción Efecto de dos variables independientes cuando actúan juntas.

Procedimientos de selección de variables Métodos para la selección de un subconjunto de variables independientes para un modelo de regresión.

Autocorrelación Correlación en los errores, que se presenta cuando los términos del error pertenecientes a puntos sucesivos de tiempo están relacionados.

Correlación serial Es lo mismo que autocorrelación.

Prueba de Durbin-Watson Prueba para determinar si existe autocorrelación de primer orden.

Fórmulas clave

Modelo lineal general

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \epsilon \qquad \textbf{(16.1)}$$

Estadístico de prueba *F* para agregar o eliminar *p* – *q* variables

$$F = \frac{\dfrac{\text{SCE}(x_1, x_2, \ldots, x_q) - \text{SCE}(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{p - q}}{\dfrac{\text{SCE}(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{n - p - 1}} \qquad \textbf{(16.13)}$$

Autocorrelación de primer orden

$$\epsilon_t = \rho\epsilon_{t-1} + z_t \qquad \textbf{(16.16)}$$

Estadístico de prueba de Durbin-Watson

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{16.17}$$

Ejercicios complementarios

29. La disminución en los precios de las impresoras láser a color, hacen de ellas una muy buena alternativa frente a las impresoras de inyección de tinta. *PC World* examinó y evaluó 10 impresoras a color. En los datos siguientes se presentan el precio, la velocidad de impresión en páginas por minuto (ppm) y la evaluación de *PC World* de estas impresoras (*PC World*, diciembre de 2005).



ColorPrinter

Fabricante y modelo	Velocidad (ppm)	Evaluación
Dell 3000cn	3.4	83
Oki Data C5200n	5.2	81
Konica Minolta MagiColor 2430DL	2.7	79
Brother HL-2700CN	3.1	78
Lexmark C522n	3.8	77
HP Color LaserJet 3600n	5.6	74
Xerox Phaser 6120n	1.6	73
Konica Minolta MagiColor 2450	1.6	71
HP Color LaserJet 2600n	2.6	70
HP Color LaserJet 2550L	1.1	61

a. Elabore un diagrama de dispersión, use como variable independiente la velocidad de impresión. ¿Un modelo de regresión simple parece apropiado?
b. Obtenga una ecuación estimada de regresión múltiple en la que las variables independientes sean x = velocidad y x^2.
c. Considere el modelo no lineal indicado por la ecuación (16.7). Use logaritmos para transformar este modelo no lineal en un modelo lineal equivalente y obtenga la ecuación estimada de regresión correspondiente. ¿Esta ecuación estimada de regresión proporciona un mejor ajuste que la ecuación estimada de regresión obtenida en el inciso b?

30. Muchos fondos internacionales ofrecen tasas más razonables que en Estados Unidos. Como los mercados internacionales suelen moverse en direcciones distintas a los mercados de Estados Unidos, las inversiones en mercados extranjeros pueden reducir el riesgo de un inversionista. En la tabla siguiente se presentan 20 fondos internacionales dando tipo de fondo (con comisión o sin comisión), coeficiente de gastos (%), seguridad (0 = la más riesgosa, 10 = la más segura) y su desempeño en un año al 10 de diciembre de 1999 (*Mutual Funds,* febrero de 2000).



MutFunds

	Tipo de fondo	Coeficiente de gastos (%)	Seguridad	Desempeño (%)
ABN AMRO Int'l Equity "Com"	Sin comisión	1.38	6.9	36
Accessor Int'l Equity "Adv"	Sin comisión	1.59	7.1	42
Artisan International	Sin comisión	1.45	6.8	72
Columbia Int'l Stock	Sin comisión	1.56	7.1	54
Concert Inv. "A" Int'l Equity	Con comisión	2.16	6.3	116
Diversified Invstr Int'l Eqty	Sin comisión	1.40	7.3	54

(*continúa*)

	Tipo de fondo	Coeficiente de gastos (%)	Seguri-dad	Desempeño (%)
Driehaus Int'l Growth	Sin comisión	1.88	6.5	92
Founders Passport	Sin comisión	1.52	7.0	86
Guardian Baillie Fifford Int'l "A"	Con comisión	1.62	7.1	37
Jamestown Int'l Equity	Sin comisión	1.56	7.1	35
Julius Baer Int'l Equity	Sin comisión	1.79	6.9	71
Aetna "I" Int'l	Sin comisión	1.35	7.3	46
Pilgrim Int'l Value "A"	Con comisión	1.80	7.1	42
Fidelity Diversified Int'l	Sin comisión	1.48	7.5	42
Putnam "A" Int'l Growth	Con comisión	1.59	6.9	55
Sit Int'l Growth	Sin comisión	1.50	6.9	49
Touchstone Int'l Equity "A"	Con comisión	1.60	7.5	35
United Int'l Growth "A"	Con comisión	1.28	7.1	47
Vontobel Int'l Equity	Sin comisión	1.50	7.0	43
Waddell & Reed Int'l Growth "B"	Con comisión	2.46	7.0	75

a. Utilice los métodos de este capítulo para obtener una ecuación estimada de regresión que sirva para estimar el desempeño de un fondo con base en los datos proporcionados.
b. ¿La ecuación estimada de regresión obtenida en el inciso a proporciona un buen ajuste? Explique.
c. Acorn International es un fondo sin comisión cuyo coeficiente de gastos es 1.12% y cuya seguridad es 7.6. Use la ecuación estimada de regresión obtenida en el inciso a para estimar el desempeño en un año de Acorn International.

31. En un estudio se investigó la relación entre el retraso en la auditoría, tiempo transcurrido desde el fin del año fiscal de una empresa hasta la fecha del informe del auditor, y variables que describen al cliente y al auditor. A continuación se presentan algunas de las variables independientes incluidas en este estudio.

Industria — Variable ficticia que toma el valor 1 si se trata de una industria y 0 si se trata de un banco, de una institución de ahorro, de préstamo o de seguros.
Pública — Variable ficticia que toma el valor 1 si la empresa es comercializada en la bolsa o es extra bursátil; 0 si no es así.
Calidad — Medida de la calidad general de los controles internos, a juicio del auditor, con una escala de cinco puntos que van desde "prácticamente ninguna" (1) hasta "excelente" (5).
Terminado — Una medida que va de 1 a 4, a juicio del auditor, donde 1 indica "todo el trabajo realizado después del fin de año" y 4 indica "la mayor parte del trabajo realizado antes del fin de año".

En una muestra de 40 empresas se obtuvieron los datos siguientes.



Retraso	Industria	Pública	Calidad	Terminado
62	0	0	3	1
45	0	1	3	3
54	0	0	2	2
71	0	1	1	2
91	0	0	1	1
62	0	0	4	4
61	0	0	3	2
69	0	1	5	2
80	0	0	1	1
52	0	0	5	3

Retraso	Industria	Pública	Calidad	Terminado
47	0	0	3	2
65	0	1	2	3
60	0	0	1	3
81	1	0	1	2
73	1	0	2	2
89	1	0	2	1
71	1	0	5	4
76	1	0	2	2
68	1	0	1	2
68	1	0	5	2
86	1	0	2	2
76	1	1	3	1
67	1	0	2	3
57	1	0	4	2
55	1	1	3	2
54	1	0	5	2
69	1	0	3	3
82	1	0	5	1
94	1	0	1	1
74	1	1	5	2
75	1	1	4	3
69	1	0	2	2
71	1	0	4	4
79	1	0	5	2
80	1	0	1	4
91	1	0	4	1
92	1	0	1	4
46	1	1	4	3
72	1	0	5	2
85	1	0	5	1

a. Obtenga una ecuación estimada de regresión con todas las variables independientes.
b. ¿La ecuación estimada de regresión obtenida en el inciso a proporciona un buen ajuste?
c. Trace un diagrama de dispersión en el que se presente la variable retraso en función de la variable terminado.
d. Con base en sus observaciones acerca de la relación entre retraso y terminado obtenga otra ecuación estimada de regresión, distinta a la dada en el inciso a, que explique la mayor proporción posible de la variabilidad de retraso.

32. Remítase a los datos del ejercicio 31. Considere un modelo en el que para predecir retraso se use únicamente industria. Emplee como nivel de significancia 0.01 y pruebe si existe alguna autocorrelación en los datos.

33. Remítase a los datos del ejercicio 31.
    a. Obtenga una ecuación estimada de regresión para predecir retraso empleando industria y calidad.
    b. Grafique los residuales obtenidos con la ecuación estimada de regresión obtenida en el inciso a en función del orden en que están presentados los datos. ¿Parece existir alguna autocorrelación en los datos? Explique.
    c. Con un nivel de significancia 0.05, pruebe si existe alguna autocorrelación en los datos.

34. Se realizó un estudio para investigar la actividad de los compradores cuando buscan y miran cosas dentro de una tienda, y de acuerdo con esto se les clasificó como inactivos, poco activos y muy activos. También se midió qué tan cómodo se sentía cada comprador en la tienda; puntuaciones más altas correspondían a mayor comodidad. Los datos siguientes provienen de este estudio. Emplee como nivel de significancia 0.05 y realice una prueba para determinar las diferencias que existen en la comodidad dentro de la tienda entre los tres tipos de compradores.



Inactivos	Poco activos	Muy activos
4	5	5
5	6	7
6	5	5
3	4	7
3	7	4
4	4	6
5	6	5
4	5	7

35. La revista *Money* publicó precios y diversos datos de los 418 vehículos más populares entre los modelos del 2003. Una de estas variables fue el valor de reventa del vehículo, expresado como porcentaje del precio de reventa sugerido por el fabricante. Estos datos se clasificaron de acuerdo con el tamaño y tipo de vehículo. En la tabla siguiente se presentan los valores de reventa de 10 automóviles pequeños elegidos aleatoriamente, de 10 automóviles medianos elegidos aleatoriamente, de 10 automóviles de lujo elegidos aleatoriamente y de 10 automóviles deportivos elegidos aleatoriamente (*Money*, marzo de 2003).



Pequeño	Mediano	De lujo	Deportivo
26	26	36	41
31	29	38	39
41	41	38	30
32	27	39	34
27	26	35	40
34	33	26	43
31	27	40	42
38	29	47	39
27	35	41	44
42	39	32	50

Use $\alpha = 0.05$ para determinar si existe alguna diferencia significativa entre los valores medios de reventa de los cuatro tipos de automóviles.

Caso problema 1 Análisis de las estadísticas de la PGA Tour



La Professional Golfers Association (PGA) lleva un registro sobre ganancias y datos de desempeño de sus miembros en la PGA Tour. En el archivo PGA Tour del disco compacto se presentan los datos de fin de año sobre el desempeño de los 125 jugadores que tuvieron los mejores ingresos en los eventos de la PGA Tour de 2005 (www.pgatour.com, 2006). Cada renglón del conjunto de datos corresponde a un jugador de la PGA Tour, y los datos han sido ordenados con base en las ganancias totales. A continuación se presenta la descripción de los datos.

Earnings	Ganancias totales en los eventos de la PGA Tour
Scoring Avg.	Puntuación promedio de un jugador en todos los eventos
Yards/Drive	Promedio de yardas por salto p. 740
Driving Acc.	Porcentaje de veces que el jugador llega a la calle con un tee shot
Greens in Reg.	Porcentaje de veces que el jugador llega al green en regulación se considera como golpe a un green en regulación, si cualquier parte de la bola está tocando la superficie y la diferencia entre el valor del par para el hoyo y el número de los movimientos para golpear el green es por lo menos 2.

Putting Avg.	Promedio de putts (toques) realizados en el green en regulación
Save Pct.	Porcentaje de veces que el jugador logra “subir y bajar” (“up and down”) cuando se encuentra en un búnker de arena al lado del green

Informe administrativo

Suponga que un representante de la PGA Tour lo contrata para analizar los datos para una presentación que se realizará en la reunión anual de la PGA Tour. Este representante le pregunta si es posible usar estos datos para determinar una medida del desempeño que sea el mejor predictor de la puntuación promedio de un jugador. Use los métodos presentados en este capítulo y en los capítulos anteriores para analizar estos datos. Formule un informe para el representante de la PGA Tour en el que resuma su análisis y en el que incluya los resultados estadísticos más importante, sus conclusiones y recomendaciones. En un apéndice presente todo el material técnico que considere adecuado.

Caso problema 2 Rendimiento de combustible en los automóviles



Cars

En todos los automóviles nuevos que se venden en Estados Unidos, viene una etiqueta sobre el consumo de combustible indicando el rendimiento en millas por galón que se espera del automóvil tanto en ciudad como en carretera. En la *Fuel Economy Guide* del Departamento de Energía de Estados Unidos se encuentra esta información para cualquier automóvil o camión. En el archivo Cars del disco compacto que se distribuye con el libro se encuentra parte de estos datos para 230 automóviles (www.fueleconomy.gov, 21 de marzo de 2003). A continuación se presenta una descripción de los datos que vienen en el disco compacto.

Class	Tipo de automóvil (compacto, mediano, grande)
Manufacturer	Empresa fabricante del automóvil
carline name	Nombre del automóvil
displ	Desplazamiento del motor en litros
cyl	Cilindros que tiene el motor (4, 6, 8)
trans	Tipo de transmisión (automática, manual)
cty	Consumo de combustible en la ciudad en millas por galón
hwy	Consumo de combustible en carretera en millas por galón

Informe administrativo

Emplee los métodos presentados en este capítulo y en los anteriores y analice este conjunto de datos. El objetivo es obtener una ecuación estimada de regresión que sirva para estimar el consumo de combustible en la ciudad y una ecuación estimada de regresión que sirva para estimar el consumo de combustible en carretera. De su análisis, presente un resumen, en el que incluya los resultados estadísticos más importantes, sus conclusiones y recomendaciones. En un apéndice incluya cualquier material técnico que considere adecuado (resultados de computadora, gráficas de residuales, etc.).

Caso problema 3 Predicción de las tasas de alumnos que llegan a titularse en las universidades

Para los administradores universitarios, el porcentaje de alumnos que ingresan a una universidad y que llegan hasta su titulación es un dato estadístico importante. Algunos de los factores que están relacionados con el porcentaje de alumnos que llegan hasta la titulación es el porcentaje de

clases en las que hay menos de 20 alumnos, el porcentaje de clases en las que hay más de 50 estudiantes, la proporción de estudiantes por facultad, el porcentaje de estudiantes que solicitan ingresar a la universidad y que son admitidos, el porcentaje de estudiantes de primer ingreso que estuvieron en el 10% más alto de sus clases de bachillerato y la reputación académica de la universidad. Para estudiar el efecto de estos factores sobre el porcentaje de alumnos que llegan a la titulación, se recolectaron datos de 48 universidades de Estados Unidos (*America's Best Colleges*, Edición del año 2000). Estos datos se encuentran en el archivo GradeRate del disco compacto. A continuación se presenta una descripción de los datos que aparecen en el disco.



GradRate

Region	Región del país en donde se encuentra la universidad
Graduation Rate	Porcentaje de estudiantes que entran a la universidad y que se titulan
% of classes under 20	Porcentaje de las clases en las que hay menos de 20 alumnos
% of classes of 50 or more	Porcentaje de las clases en las que hay más de 50 alumnos
Student-Faculty Ratio	Cociente del número de estudiantes inscritos dividido entre el número de profesores
Acceptance rate	Porcentaje de estudiantes que solicitan inscripción a la universidad y que son aceptados
1st-Year students in top 10% of HS class	De los estudiantes admitidos a la universidad, porcentaje que estuvo en el 10% más alto de sus clases de bachillerato
Academic Reputation Score	Una medida de la reputación de la universidad determinada mediante una revisión a los administradores en otras universidades: medida en una escala del 1 (marginal) al 5 (distinguida)

Informe administrativo

Use los métodos presentados en este capítulo y en los capítulos anteriores para analizar este conjunto de datos. Presente un resumen de su análisis en el que dé los principales resultados estadísticos, sus conclusiones y recomendaciones, en un informe administrativo. En un apéndice presente cualquier material técnico (resultados de computadora, gráficas de residuales, etc.) que considere adecuados.

Apéndice 16.1 Procedimientos de selección de variables con Minitab

En la sección 16.4 se vio el uso de los procedimientos de selección de variables para la solución de problemas de regresión múltiple. En la figura 16.16 se mostraron los resultados que da la regresión por pasos de Minitab aplicada a los datos de Cravens y en la figura 16.17 los resultados que da el procedimiento de los mejores subconjuntos de Minitab. En este apéndice se describen los pasos necesarios para obtener los resultados que se muestran en esas dos figuras, así como los pasos que se requieren en los procedimientos de selección hacia adelante y eliminación hacia atrás. Primero, en una hoja de cálculo de Minitab se ingresan los datos de la tabla 16.5. Los valores de Sales, Time, Poten, AdvExp, Share, Change, Accounts y Rating se ingresan en las columnas C1-C9 de la hoja de cálculo de Minitab.

Uso del procedimiento por pasos de Minitab

Mediante los pasos siguientes se obtienen los resultados de la regresión por pasos de Minitab para los datos de Cravens.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Seleccionar el menú **Regression**
Paso 3. Elegir **Stepwise**

Paso 4. Cuando aparezca el cuadro de diálogo **Stepwise Regression:**
Ingresar Sales en el cuadro de diálogo **Response**
Ingresar Time, Poten, AdvExp, Share, Change, Accounts y Rating en el cuadro **Predictors**
Seleccionar el botón **Methods**

Paso 5. Cuando aparezca el cuadro de diálogo **Stepwise Method**:
Seleccionar **Stepwise** (**forward and backward**)
Ingresar 0.05 en el cuadro **Alpha to enter**
Ingresar 0.05 en el cuadro **Alpha to remove**
Clic en **OK**

Paso 6. Cuando aparezca el cuadro de diálogo **Stepwise Regression:**
Clic en **OK**

Uso del procedimiento de selección hacia adelante de Minitab

Para usar el procedimiento de selección hacia adelante de Minitab, sólo hay que modificar el paso 5 del procedimiento de regresión por pasos, como se indica a continuación:

Paso 5. Cuando aparezca el cuadro de diálogo **Stepwise-Methods:**
Seleccionar **Forward Selection**
Ingresar 0.05 en el cuadro de diálogo **Alpha to enter**
Clic en **OK**

Uso del procedimiento de eliminación hacia atrás de Minitab

Para usar el procedimiento de eliminación hacia atrás de Minitab, sólo hay que modificar el paso 5 del procedimiento de regresión por pasos, como se indica a continuación:

Paso 5. Cuando aparezca el cuadro de diálogo **Stepwise-Methods:**
Seleccionar **Backward elimination**
Ingresar 0.05 en el cuadro de diálogo **Alpha to remove**
Clic en **OK**

Uso del procedimiento de los mejores subconjuntos de Minitab

Mediante los pasos siguientes se obtienen los resultados para los datos de Cravens que da la regresión de los mejores subconjuntos de Minitab.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Seleccionar el menú **Regression**
Paso 3. Elegir **Best Subsets**
Paso 4. Cuando aparezca el cuadro de diálogo **Best Subsets Regression**
Ingresar Sales en el cuadro **Response**
Ingresar Time, Poten, AdvExp, Share, Change, Accounts y Rating en el cuadro **Predictors**
Clic en **OK**



CAPÍTULO 17

Números índice

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

DEPARTAMENTO DEL TRABAJO DE ESTADOS UNIDOS, DEPARTAMENTO DE ESTADÍSTICA LABORAL
WASHINGTON, D.C.

El Departamento del Trabajo de Estados Unidos, a través de su Departamento de Estadística Laboral, recopila y hace circular índices que sirven como indicadores de la actividad comercial y económica de Estados Unidos. Por ejemplo, este departamento recopila y publica el índice de precios al consumidor, el índice de precios al productor, y otras estadísticas sobre horas y ganancias promedio de varios tipos de trabajadores. Quizás el índice más citado de los elaborados por el Departamento de Estadística Laboral es el índice de precios al consumidor. Este índice suele emplearse como una medida de la inflación.

En febrero de 2006, el Departamento de Estadística Laboral informó que el índice de precios al consumidor (IPC) había tenido un aumento de 0.2% desde enero. Este incremento se presentaba después de un incremento de 0.7% en enero y parecía indicar una desaceleración de la tasa de inflación. El Departamento de Estadística Laboral también informó que la "tasa central" de la inflación había sido de sólo 0.1% en febrero. En la "tasa central" se excluyen del IPC los componentes volátiles que son alimentos y combustibles (energía), y se le suele considerar como un mejor indicador de las presiones inflacionarias. El costo de los combustibles (energía) aumentó 5% en enero y fue la principal razón de que el IPC tuviera un incremento de 0.7% en ese mes.

El Departamento de Estadística Laboral informó que el índice de precios al productor (IPP) había disminuido 1.4% en febrero de 2006. El IPP mide la variación de los precios en los mercados mayoristas y suele considerarse como el principal indicador de las variaciones en el índice de precios al consumidor. Gran parte de esa disminución se debió a la disminución de los precios de los bienes de energía. Al eliminar alimentos y combustibles (energía), el IPP en realidad aumentó en el mes de febrero.

En este capítulo se verá cómo se calculan diversos índices, como los índices de precios al consumidor y al productor y cómo deben interpretarse.



El precio de la gasolina es un componente del índice del precio del consumidor. © AP Photo/Jeff Chiu.

Todos los meses, el gobierno de Estados Unidos publica diversos índices con objeto de ayudar a las personas a entender las condiciones económicas y comerciales vigentes. El más conocido y citado de estos índices es, probablemente, el índice de precios al consumidor (IPC). Como su nombre lo indica, el IPC es un indicador de lo que ocurre con los precios que pagan los consumidores por los artículos que compran. En concreto, el IPC mide las variaciones en los precios a lo largo de un lapso. Al partir de un punto determinado o *periodo base* y de su índice correspondiente, que es 100, el IPC sirve para comparar los precios al consumidor del periodo actual con los del periodo base. Por ejemplo, si el IPC es 125, esto significa que los precios, como un todo, al consumidor son aproximadamente 25% más altos que los precios de los mismos artículos en el periodo base. Aun cuando relativamente pocas personas saben exactamente lo que significan estos números, acerca del IPC todo mundo sabe lo suficiente para entender que su aumento significa precios más altos.

A pesar de que el IPC es tal vez el índice más conocido, hay muchos otros índices gubernamentales y del sector privado que ayudan a medir y a entender las condiciones económicas de un periodo, en comparación con las de otro periodo. El objetivo de este capítulo es describir los tipos de índices más comunes. Para empezar se construirán algunos números índice sencillos para poder entender mejor cómo se calculan.



Precios relativos

TABLA 17.1

COSTO DE LA GASOLINA NORMAL

Año	Precio por galón ($)
1990	1.30
1991	1.10
1992	1.09
1993	1.07
1994	1.08
1995	1.11
1996	1.22
1997	1.20
1998	1.03
1999	1.14
2000	1.48
2001	1.42
2002	1.34
2003	1.56
2004	1.85
2005	2.27

Fuente: U.S. Energy Information Administration.

La forma más sencilla de un índice de precios permite hacer una comparación entre el precio actual por unidad de un artículo y el precio por unidad del mismo artículo en el periodo base. Por ejemplo, en la tabla 17.1 se presentan los precios de un galón de gasolina desde 1990 hasta 2005. Para facilitar la comparación con otros años, se convierte el precio actual por galón a un **precio relativo** que expresa el precio unitario en cada periodo como un porcentaje del precio unitario en el periodo base.

$$\text{Precio relativo en el periodo } t = \frac{\text{Precio en el periodo } t}{\text{Precio en el periodo base}}(100) \tag{17.1}$$

Con 1990 como año base para los precios de la gasolina de la tabla 17.1, se pueden calcular los precios relativos de un galón de gasolina normal desde 1990 hasta 2005. En la tabla 17.2 se presentan estos precios relativos. Note lo fácil que es: el precio de cualquier año se compara con el precio del año base y el resultado es el precio relativo. Por ejemplo, el precio relativo de 1995 es 85.4, lo que indica que en 1995 la gasolina costaba 14.6% menos que en el año base 1990. De manera similar, en el año 2002 el precio relativo fue 103.1, lo que indica que en el 2002 hubo un incremento de 3.1% en el costo de la gasolina en comparación con el año base 1990. Los precios relativos, como los de la gasolina normal, son muy útiles para entender e interpretar cambios en las condiciones comerciales y económicas a través del tiempo.



Índices de precios agregados

Aunque los precios relativos sirven para identificar las variaciones a lo largo del tiempo en los precios de artículos individuales, suele tenerse más interés en las variaciones en el precio de un conjunto de artículos considerados como un todo. Por ejemplo, si se desea tener un índice que mida la variación del costo general de la vida a lo largo del tiempo, el índice deberá basarse en la variación de los precios de diversos artículos como alimentos, vivienda, vestido, transporte, asistencia médica, etc. Un **índice de precios agregados** tiene como propósito medir la variación combinada de un grupo de artículos.

TABLA 17.2

PRECIOS RELATIVOS DE UN GALÓN DE GASOLINA NORMAL

Año	Precio relativo (Base 1990)
1990	(1.30/1.30)100 = 100.0
1991	(1.10/1.30)100 = 84.6
1992	(1.09/1.30)100 = 83.8
1993	(1.07/1.30)100 = 82.3
1994	(1.08/1.30)100 = 83.1
1995	(1.11/1.30)100 = 85.4
1996	(1.22/1.30)100 = 93.8
1997	(1.20/1.30)100 = 92.3
1998	(1.03/1.30)100 = 79.2
1999	(1.14/1.30)100 = 87.7
2000	(1.48/1.30)100 = 113.8
2001	(1.42/1.30)100 = 109.2
2002	(1.34/1.30)100 = 103.1
2003	(1.56/1.30)100 = 120.0
2004	(1.85/1.30)100 = 142.3
2005	(2.27/1.30)100 = 174.6

Considere, por ejemplo, un índice de precios agregados de un grupo de artículos catalogados como gastos por el uso de un automóvil. Para ejemplificar, los artículos comprendidos en este grupo se limitarán a gasolina, aceite, neumáticos y gastos de seguro.

En la tabla 17.3 se presentan los datos, para este índice de gastos, correspondientes a los años 1990 y 2005. Con 1990 como periodo base, un índice de precios agregados para estos cuatro componentes proporcionará una medida de la variación en el periodo de 1990 a 2005 en los gastos por el uso de un automóvil.

Un índice agregado no ponderado se obtiene al sumar los precios unitarios en el año de interés (en este caso, 2005) y dividir esta suma entre la suma de los precios unitarios en el año base (1990). Sean

$$P_{it} = \text{precio unitario del artículo } i \text{ en el periodo } t$$
$$P_{i0} = \text{precio unitario del artículo } i \text{ en el periodo base}$$

Un índice agregado no ponderado del periodo t, que se denota I_t, está dado por

$$I_t = \frac{\Sigma P_{it}}{\Sigma P_{i0}}(100) \tag{17.2}$$

donde las sumas son de todos los artículos del grupo.

TABLA 17.3 DATOS PARA EL ÍNDICE DE GASTOS POR EL USO DE UN AUTOMÓVIL

Artículo	Precio unitario ($) 1990	2005
Galón de gasolina	1.30	2.27
Cuarto de galón de aceite	2.10	3.50
Neumáticos	130.00	170.00
Seguro	820.00	939.00

Un índice agregado no ponderado de los gastos por el uso de un automóvil en 2005 ($t = 2005$) está dado por

$$I_{2005} = \frac{2.27 + 3.50 + 170.00 + 939.00}{1.30 + 2.10 + 130.00 + 820.00}(100)$$

$$= \frac{1114.77}{953.4}(100) = 117$$

De acuerdo con este índice de precios agregados no ponderados, se concluye que, en el periodo de 1990 a 2005, el precio de los gastos por el uso de un automóvil ha aumentado 17%. Pero hay que notar que en este índice de precios no ponderados por los gastos del uso de un automóvil influyen más los artículos cuyo precio unitario es elevado. Por tanto, artículos que tienen precios unitarios bajos, como gasolina y aceite, están dominados por los artículos de precio unitario alto, como neumáticos y seguro. La influencia que tienen los precios de los neumáticos y del seguro sobre el índice agregado no ponderado de los gastos por el uso de un automóvil es muy grande.

Si la cantidad usada de cada uno de los artículos es la misma, con un índice no ponderado se obtienen los mismos resultados que con un índice ponderado. Sin embargo, en la práctica, las cantidades usadas rara vez son iguales.

Debido a la sensibilidad que muestran los índices no ponderados de uno o varios artículos de precio más elevado, este tipo de índice agregado no es muy usado. Cuando las cantidades usadas son diferentes, con un índice de precios agregados ponderados se obtiene una mejor comparación.

La filosofía que hay detrás del **índice de los precios agregados ponderados** es que cada artículo del grupo debe ser ponderado de acuerdo con su importancia. Por tanto, se necesita una medida de la cantidad de uso de cada artículo del grupo. En la tabla 17.4 se presenta la información del uso anual de cada artículo que se debe tener en cuenta en el uso de un automóvil, con base en el uso estándar de un automóvil mediano que recorre alrededor de 15 000 millas anuales. Los ponderadores de las cantidades mostrados indican el uso anual esperado en estas condiciones.

TABLA 17.4 INFORMACIÓN SOBRE USO ANUAL PARA EL ÍNDICE DE USO DE UN AUTOMÓVIL

Artículo	Ponderador de la cantidad*
Galones de gasolina	1000
Cuartos de galón de aceite	15
Neumáticos	2
Seguro	1

* Con base en 15 000 millas por año. La vida de uso de los neumáticos es de 30 000 millas.

Sea Q_i = cantidad usada del artículo i. El índice de precio agregado ponderado del periodo t está dado por

$$I_t = \frac{\Sigma P_{it} Q_i}{\Sigma P_{i0} Q_i}(100) \tag{17.3}$$

donde las sumas son de todos los artículos del grupo. Aplicado a los gastos por el uso de un automóvil, el índice de precios agregados ponderados se obtiene al dividir los costos del uso del automóvil en 2005 entre los costos de uso del automóvil en 1990.

Si $t = 2005$, con los ponderadores de la tabla 17.4 se obtiene el siguiente índice de precios agregados ponderados para los gastos por el uso de un automóvil en 2005.

$$I_{2005} = \frac{2.27(1000) + 3.50(15) + 170.00(2) + 939.00(1)}{1.30(1000) + 2.10(15) + 130.00(2) + 820.00(1)}(100)$$

$$= \frac{3601.5}{2411.5}(100) = 149$$

De acuerdo con este índice de precios agregados ponderados, se puede concluir que durante el periodo 1990 a 2005, el precio de los gastos por el uso de un automóvil ha aumentado 49%.

Es claro que en comparación con el índice agregado no ponderado, el índice ponderado da una indicación más precisa de la variación que ha habido de 1990 a 2005, en el precio de los gastos por el uso de un automóvil. Al tomar en cuenta la cantidad usada de gasolina, se contrarresta el pequeño aumento porcentual de los costos del seguro. Con el índice ponderado se obtiene un aumento mayor, en los gastos por el uso de un automóvil, que con el índice no ponderado. En general, para establecer un índice de precios para un grupo de artículos se prefieren los índices agregados ponderados con las cantidades de uso como cargas.

Observe que en la fórmula (17.3) para el índice de precios agregados ponderados, el término correspondiente a la cantidad Q_i no tiene un segundo subíndice que indique el tiempo. La razón es que las cantidades Q_i son fijas, no varían con el tiempo como lo hacen los precios. Estos ponderados fijos o cantidades los especifica la persona que diseña el índice al emplear las cantidades que considere representativas del uso estándar. Una vez establecidas estas cantidades se mantienen constantes o fijas para todos los periodos que se use el índice. Para obtener índices de otros años que no sean el 2005, es necesario recolectar nuevos datos P_{it}, pero las cantidades ponderadoras Q_i permanecen constantes.

En un caso especial del índice agregado de ponderadores fijos, las cantidades se determinan de acuerdo con el uso en el año base. En este caso se escribe $Q_i = Q_{i0}$, donde el cero que se emplea como subíndice indica el ponderador de la cantidad del año base; la fórmula (17.3) se convierte en

$$I_t = \frac{\Sigma P_{it} Q_{i0}}{\Sigma P_{i0} Q_{i0}}(100) \tag{17.4}$$

Cuando los ponderadores fijos de las cantidades se determinan a partir de los usos en el año base, al índice agregado ponderado se le conoce como **índice de Laspeyres**.

Otra posibilidad para determinar los ponderadores de las cantidades es ir modificando estas cantidades en cada periodo. En este caso, cada año para el que se calcula el índice se determina Q_{it}. El índice agregado ponderado del periodo t con estos ponderadores de las cantidades está dado por

$$I_t = \frac{\Sigma P_{it} Q_{it}}{\Sigma P_{i0} Q_{it}}(100) \tag{17.5}$$

Observe que tanto en el periodo base (periodo 0) como en el periodo t se usan los mismos ponderadores de las cantidades. Pero las ponderaciones están basadas en el uso en el periodo t, no en el periodo base. A este índice agregado ponderado se le conoce como **índice de Paasche**. Este índice tiene la ventaja de estar basado en los estándares de uso actuales. Sin embargo, este método de cálculo de un índice agregado ponderado tiene dos desventajas: las cantidades del uso Q_{it} tienen que ser determinadas cada año, sumando al tiempo y al costo de la recolección de datos, y cada año hay que volver a calcular los números índice de los años anteriores para que pueda verse el efecto de los nuevos pesos. Debido a esta desventaja, el índice más usado es el índice de Laspeyres. El índice de los gastos por el uso de un automóvil se calculó con las cantidades del periodo base; por tanto, es un índice de Laspeyres. Si se hubieran usado las cifras correspondientes al 2005, hubiera sido un índice de Paasche. En realidad, debido a que los automóviles han ido teniendo un consumo de gasolina más eficiente, el uso de gasolina ha disminuido y con el índice de Paasche se obtiene una cifra distinta que con el índice de Laspeyres.

Ejercicios

Métodos



1. En la tabla siguiente se presentan precios y cantidades usadas de dos productos correspondientes a 2004 y a 2006.

	Cantidad		Precio unitario ($)	
Artículo	**2004**	**2006**	**2004**	**2006**
A	1500	1800	7.50	7.75
B	2	1	630.00	1500.00

a. Calcule los precios relativos de cada artículo en el 2006 con 2004 como periodo base.
b. Calcule un índice de precios agregados no ponderados de estos dos artículos en 2006, use 2004 como periodo base.
c. Calcule un índice de precios agregados ponderados de estos dos artículos con el método de Laspeyres.
d. Calcule un índice de precios agregados ponderados de estos dos artículos con el método de Paasche.

2. En 2006 un artículo cuyo precio relativo es 132 cuesta $10.75. Su año base es 1992.
   a. ¿En qué porcentaje aumentó o disminuyó el costo de este artículo en este lapso de 14 años?
   b. ¿Cuánto costaba este artículo en 1992?

Aplicaciones



3. Un fabricante tiene tres proveedores de un determinado componente; los tres proveedores difieren en calidad y cantidad que suministran. En la tabla siguiente se presentan los datos correspondientes a los años 2004 y 2006

		Precio unitario ($)	
Proveedor	**Cantidad (2004)**	**2004**	**2006**
A	150	5.45	6.00
B	200	5.60	5.95
C	120	5.50	6.20

   a. Calcule, por separado, los precios relativos de cada proveedor. Compare el incremento de precios de los proveedores en este lapso de dos años.
   b. Calcule un índice de precios agregados no ponderados de los componentes en el 2006.
   c. Calcule un índice de precios agregados ponderados de los componentes en el 2006. ¿Qué significado tiene este índice para el fabricante?

4. La empresa R&B Beverages, Inc., tiene toda una línea de cervezas, vinos y refrescos que distribuye a través de minoristas en Iowa central. En la tabla siguiente se presentan precios unitarios en 2003 y en 2006 y cantidades vendidas (cajas) en 2003.

		Precio unitario ($)	
Artículo	**Cantidades 2003 (cajas)**	**2003**	**2006**
Cerveza	35 000	16.25	17.50
Vino	5 000	64.00	100.00
Refresco	60 000	7.00	8.00

   Calcule el índice agregado ponderado de las ventas de R&B en 2006, emplee 2003 como periodo base.

5. En el método LIFO de evaluación de inventario se debe establecer un índice de precios del inventario para fines de impuestos. Los pesos se basan en los niveles de inventario de fin de año. Use el precio unitario de principio de año como precio del periodo base y obtenga un índice agregado ponderado del valor total de inventario a fin de año. ¿Qué tipo de índice de precios agregados ponderados debe emplearse para la evaluación LIFO de inventario?

		Precio unitario ($)	
Producto	**Inventario final**	**Principio**	**Fin**
A	500	0.15	0.19
B	50	1.60	1.80
C	100	4.50	4.20
D	40	12.00	13.20



17.3 Cálculo de un índice de precios agregados a partir de precios relativos

En la sección 17.1 se definió el concepto de precio relativo y se mostró cómo calcular un precio relativo a partir del precio unitario en el periodo actual y del precio unitario en el periodo base. Ahora se quiere mostrar cómo calcular directamente, a partir de la información sobre el precio relativo de cada artículo, índices de precios relativos como los obtenidos en la sección 17.2. Debido al uso limitado de los índices no ponderados, se restringirá la atención a los índices de precios agregados ponderados. Se vuelve a los índices de los gastos por el uso de un automóvil de la sección anterior. La información necesaria acerca de los cuatro artículos se encuentra en la tabla 17.5.

Hay que comprobar que los precios y las cantidades se den en las mismas unidades. Por ejemplo, si los precios son precios por caja, las cantidades deben darse en cantidad de cajas y no, por ejemplo, en cantidad de las unidades que vienen en las cajas.

Sea w_i la ponderación correspondiente al precio relativo del artículo i. La expresión general para obtener un promedio ponderado de precios relativos es la dada por

$$I_t = \frac{\sum \frac{P_{it}}{P_{i0}}(100)w_i}{\Sigma w_i} \qquad \textbf{(17.6)}$$

Elegir de manera adecuada las ponderaciones de la ecuación (17.6) permitirá calcular un índice de precios agregados ponderados a partir de los precios relativos. Las ponderaciones adecuadas son las que se obtienen al multiplicar el precio del periodo base por la cantidad usada (la cantidad de uso).

$$w_i = P_{i0}Q_i \qquad \textbf{(17.7)}$$

Sustituyendo $w_i = P_{i0}Q_i$ en la ecuación (17.6) se obtiene la ecuación siguiente que da un índice de precios relativos ponderados.

$$I_t = \frac{\sum \frac{P_{it}}{P_{i0}}(100)(P_{i0}Q_i)}{\Sigma P_{i0}Q_i} \qquad \textbf{(17.8)}$$

Como en el numerador se cancelan los términos P_{i0}, se obtiene la siguiente expresión equivalente para el índice de precios relativos ponderados

$$I_t = \frac{\Sigma P_{it}Q_i}{\Sigma P_{i0}Q_i}(100)$$

Como se ve, el índice de precios relativos ponderados en el que $w_i = P_{i0}Q_i$ proporciona un índice idéntico al índice agregado ponderado al que se obtiene con la ecuación (17.3) de la sección

TABLA 17.5 PRECIOS RELATIVOS PARA EL ÍNDICE DE GASTOS POR EL USO DE UN AUTOMÓVIL

	Precio unitario ($)			
Artículo	**1990 (P_0)**	**2005 (P_t)**	**Precio relativo (P_t/P_0)100**	**Uso anual**
Galón de gasolina	1.30	2.27	174.6	1000
Cuarto de galón de aceite	2.10	3.50	166.7	15
Neumáticos	130.00	170.00	130.8	2
Seguro	820.00	939.00	114.5	1

TABLA 17.6 ÍNDICE DE GASTOS POR EL USO DE UN AUTOMÓVIL (1990-2005) CON BASE EN LOS PRECIOS RELATIVOS PONDERADOS

Artículo	Precios relativos $(P_{it}/P_{i0})(100)$	Precio base ($) P_{i0}	Cantidad Q_i	Carga $w_i = P_{i0}Q_i$	Precios relativos ponderados $(P_{it}/P_{i0})(100)w_i$
Gasolina	174.6	1.30	1000	1300.00	226 980.00
Aceite	166.7	2.10	15	31.50	5 251.05
Neumáticos	130.8	130.00	2	260.00	34 008.00
Seguro	114.5	820.00	1	820.00	93 890.00
			Totales	2411.50	360 129.05

$$I_{2005} = \frac{360,129.05}{2411.50} = 149$$

17.2. Al usar en la ecuación (17.7) las cantidades del periodo base (es decir, $Q_i = Q_{i0}$) se obtiene el índice de Laspeyres. En la ecuación (17.7) si usan las cantidades del periodo actual (es decir, $Q_i = Q_{it}$) se obtiene el índice de Paasche.

Para los datos de los gastos por el uso de un automóvil, si usa los precios relativos de la tabla 17.5 y la ecuación (17.6) calcula el promedio ponderado de los precios relativos. En la tabla 17.6 se dan los resultados que se obtienen al usar las ponderaciones especificadas por la ecuación (17.7). El número índice, 149, indica que ha habido un aumento de 49% en los gastos por el uso de un automóvil, que es el mismo aumento que se encontró con el índice agregado ponderado de la sección 17.2.

Ejercicios

Métodos



6. En la tabla siguiente se dan los precios relativos de tres artículos, así como sus precios y uso en el periodo base. Calcule un índice de precios agregados ponderados para el periodo actual.

		Periodo base	
Artículo	**Precio relativo**	**Precio**	**Uso**
A	150	22.00	20
B	90	5.00	50
C	120	14.00	40

Aplicaciones



7. La empresa Mitchell Chemical fabrica un producto químico para la industria que es una mezcla de tres ingredientes químicos. A continuación se presentan los costos al comienzo del año, los costos al final del año y la proporción de cada ingrediente en la mezcla.

	Costo por libra ($)		Cantidad (libras)
Ingrediente	**Comienzo**	**Final**	**por cada 100 libras del producto**
A	2.50	3.95	25
B	8.75	9.90	15
C	.99	.95	60

a. Calcule los precios relativos de cada uno de estos tres ingredientes.
b. Calcule el promedio ponderado de los precios relativos para obtener el índice de costo anual de las materias primas usadas en este producto. ¿Qué interpretación da al valor de este índice?

8. Un portafolio de inversiones consta de cuatro acciones. En la tabla siguiente se da el precio de compra, el precio actual y la cantidad de cada una de las acciones.

Acción	Precio de compra/ acción ($)	Precio actual/acción ($)	Cantidad de acciones
Holiday Trans	15.50	17.00	500
NY Electric	18.50	20.25	200
KY Gas	26.75	26.00	500
PQ Soaps	42.25	45.50	300

Obtenga un promedio ponderado de los precios relativos como índice del desempeño del portafolio hasta la fecha. Interprete este índice de precios.

9. Calcule los precios relativos de los productos de R&B del ejercicio 4. Utilice un promedio ponderado de los precios relativos para demostrar que con este método se obtiene el mismo índice que con el método agregado ponderado.



Algunos índices de precios importantes

Se han descrito los procedimientos que se usan para calcular índices de precios de artículos o de grupos de artículos. Ahora se verán algunos índices de precios que son indicadores importantes de la situación comercial y económica. Se considerarán el índice de precios al consumidor, el índice de precios al productor y los promedios Dow Jones.

Índice de precios al consumidor

El IPC incluye gastos en servicios (por ejemplo, gastos en médicos y dentistas) y todos los impuestos debidos a la compra y el uso de un artículo.

El **índice de precios al consumidor (IPC)**, que es publicado mensualmente por el Departamento de Estadística Laboral de Estados Unidos, es la medida principal del costo de la vida en Estados Unidos. El conjunto de artículos que se usa para elaborar este índice consta de una *canasta de mercado* de 400 artículos que comprende alimentos, vivienda, vestido, transporte y medicamentos. El IPC es un índice de precios agregados ponderados que tiene pesos fijos.* Las ponderaciones que se aplican a cada artículo en la canasta de mercado se obtienen mediante un estudio de uso entre todas las familias de Estados Unidos.

En febrero de 2006, el IPC, calculado con el índice base de 100 de 1982-1984, fue 198.7. Esta cifra indica que el costo de la canasta de mercado de bienes y servicios ha aumentado 98.7% desde el periodo base 1982-1984. En la figura 17.1 se muestra una serie de IPC de 50 años, desde 1950 hasta 2000. Observe cómo el IPC refleja la conducta inflacionaria de la economía a fines de los años setenta y comienzo de los ochenta.

Índice de precios al productor

El IPP tiene por objeto medir la variación en los precios de los bienes internos; no incluye las importaciones.

El **Índice de precios al productor (IPP)**, que también es publicado mensualmente por el Departamento de Estadística Laboral de Estados Unidos, mide las variaciones mensuales de los precios en los mercados primarios de Estados Unidos. El IPP está basado en los precios de la primera operación en cada producto de los mercados no minoristas. Todos los artículos que se

* En realidad, el Departamento de Estadística Laboral publica dos índices de precio al consumidor: uno para los consumidores urbanos (IPC-U) y otro índice de precios al consumidor ajustado para los trabajadores asalariados y de oficina (IPC-W). El IPC-U es uno de los índices más citados y se publica de manera regular en *The Wall Street Journal*.

FIGURA 17.1 ÍNDICE DE PRECIOS AL CONSUMIDOR, 1950-2000 (BASE 1982-1984 = 100)

venden en operaciones comerciales de estos mercados están representados. Abarca materia prima, productos manufacturados y productos procesados en cualquiera de los niveles de procesamiento de los productos de las industrias clasificadas como manufacturera, agrícola, forestal, de pesca, minería, gas y electricidad y servicios públicos. Este índice sirve, entre otras cosas, como el principal indicador de la tendencia futura en los precios al consumidor y en el costo de la vida. El aumento del IPP refleja que ha habido un aumento en los precios al productor, lo que al final repercutirá en precios más elevados para el consumidor.

Los pesos para los diversos artículos en el IPP están basados en el valor de los pedidos de mercancías. El promedio ponderado de los precios relativos se calcula usando el método de Laspeyres. En febrero de 2006 el IPP fue 157.8 con 100 para el año 1982.

Promedios Dow Jones

Los **promedios Dow Jones** son índices que sirven como indicadores de las tendencias de los precios y de los movimientos de acciones ordinarias. El más conocido de los índices Dow Jones es el índice industrial Dow Jones (DJIA, por sus siglas en inglés), el cual está basado en los precios de las acciones ordinarias de 30 empresas grandes; es la suma de los precios de estas acciones ordinarias dividida entre un número, que se corrige de tiempo en tiempo para ajustarlo a las escisiones de las acciones o fusiones de las empresas que participan en el índice. A diferencia de los demás índices de precios estudiados, este índice no se expresa como un porcentaje de los precios del año base. En la tabla 17.7 se enumeran las empresas usadas en febrero de 2006 para calcular el DJIA.

Charles Henry Dow publicó su primer índice accionario el 3 de julio de 1884, en el Customer's Afternoon Letter. Este primer índice comprendía 11 acciones, nueve de las cuales eran ferroviarias. Un promedio comparable con el DJIA fue publicado por primera vez el 1 de octubre de 1928.

Otros promedios Dow Jones se calculan con 20 acciones del transporte o con 15 acciones de empresas de servicio público Los promedios Dow Jones se calculan y se publican diariamente en *The Wall Street Journal* y en otras publicaciones financieras.

TABLA 17.7 LAS 30 EMPRESAS USADAS EN EL ÍNDICE INDUSTRIAL DOW JONES (MARZO 2006)

Alcoa	DuPont	J. P. Morgan Chase
Altria Group	Exxon Mobil	McDonald's
AIG	General Electric	Merck
American Express	General Motors	Microsoft
AT&T	Hewlett-Packard	Minnesota Mining
Boeing	Home Depot	Pfizer
Caterpillar	Honeywell Int'l	Procter & Gamble
Citigroup	IBM	United Technologies
Coca-Cola	Intel	Verizon
Disney	Johnson & Johnson	Wal-Mart Stores

Fuente: Barron's, 20 de marzo 2006.



17.5 Deflactar una serie mediante índices de precios

Las series se deflactan para eliminar el efecto de la inflación.

Muchas series comerciales y económicas medidas a lo largo del tiempo, como ventas de empresas, ventas industriales e inventarios, miden su monto en dólares. Estas series suelen mostrar a lo largo del tiempo un patrón de aumento de crecimiento, lo que se considera como una indicación del aumento en el volumen físico relacionado con estas actividades. Por ejemplo, un aumento de 10% en el monto en dólares de un inventario puede ser interpretado como que el inventario físico es 10% mayor. Tales interpretaciones pueden ser erróneas cuando una serie de tiempo se mide en dólares y el monto total en dólares, es combinación tanto de cambios en los precios como en las cantidades. Por tanto, en aquellos periodos en que las variaciones en los precios son significativas, las variaciones en el monto en dólares pueden no corresponder a variaciones en la cantidad, a menos que la serie de tiempo pueda ser ajustada para eliminar los efectos de las variaciones en los precios.

Por ejemplo, desde 1976 hasta 1980, la cantidad total de gastos en la industria de la construcción aumentó aproximadamente 75%. Esta cifra sugiere un excelente crecimiento en la actividad de la construcción. Pero, los precios en la construcción aumentaban en ese momento tan rápido —o algunas veces aún más— como esta tasa de 75%. En efecto, aunque los gastos totales en la construcción aumentaban, la actividad en la construcción permanecía relativamente constante o, en casos como la construcción de casas nuevas, incluso disminuía. Para interpretar correctamente la actividad en la construcción durante este periodo 1976-1980, es necesario ajustar la serie de los gastos totales mediante un índice de precios para eliminar el efecto del aumento de precios. Siempre que se elimina el efecto del aumento de precios de una serie de tiempo, se dice que se está *deflactando la serie.*

En relación con el ingreso de las personas y con los salarios se suele oír discusiones acerca de "salarios reales" o "poder adquisitivo" del salario. Estos conceptos están relacionados con la noción de deflactar un índice de salario por hora. Por ejemplo, en la figura 17.2 se muestra el patrón seguido por los salarios por hora de los trabajadores de la industria en el periodo 1998-2002. Se observa una tendencia de aumento en los salarios, de $12.78 por hora a $14.77 por hora. ¿Los trabajadores de la industria estarán contentos con este aumento de los salarios por hora? La respuesta depende de lo que ocurra con el poder adquisitivo de sus salarios. Si se puede comparar el poder adquisitivo del salario de $12.78 por hora de 1998 con el poder adquisitivo del salario de $14.77 de 2002, se estará en mejores condiciones para juzgar la mejora relativa del salario.

En la tabla 17.8 se presentan tanto salarios por hora como el IPC desde 1998 hasta 2002.Con estos datos se mostrará cómo usar el IPC para deflactar el índice de los salarios por hora. La serie deflactada se obtiene al dividir el salario por hora de cada año entre el valor correspondiente

FIGURA 17.2 SALARIOS POR HORA DE LOS TRABAJADORES DEL SECTOR PRODUCCIÓN

del IPC y multiplicarlo por 100. En la tabla 17.9 se da el índice deflactado de los salarios por hora de los trabajadores de la industria; en la figura 17.3 se presenta una gráfica que muestra el salario deflactado o real.

Los salarios reales son una mejor medida del poder de compra que los salarios nominales. Muchos contratos sindicales piden que los salarios se ajusten de acuerdo con los cambios en el costo de la vida.

¿Qué indica la serie deflactada de salarios acerca de los salarios reales o poder adquisitivo de los trabajadores durante el periodo 1998-2002? En términos de dólares, en el periodo base (1982-1984 = 100), el salario por hora no aumentó mucho. Una vez eliminado el efecto inflacionario, se ve que el poder adquisitivo de los trabajadores no aumentó mucho. Este efecto se ve en la figura 17.3. Por tanto, la ventaja de usar índices de precios para deflactar una serie es que se obtiene una imagen más clara de los cambios reales, en dólares, ocurridos.

Este proceso de deflactar una serie a lo largo del tiempo tiene una aplicación importante en el cálculo del producto interno bruto (PIB). El PIB es el valor total de todos los bienes y servi-

TABLA 17.8 SALARIOS POR HORA PARA LOS TRABAJADORES DE LA INDUSTRIA E ÍNDICES DE PRECIOS AL CONSUMIDOR, 1998-2002

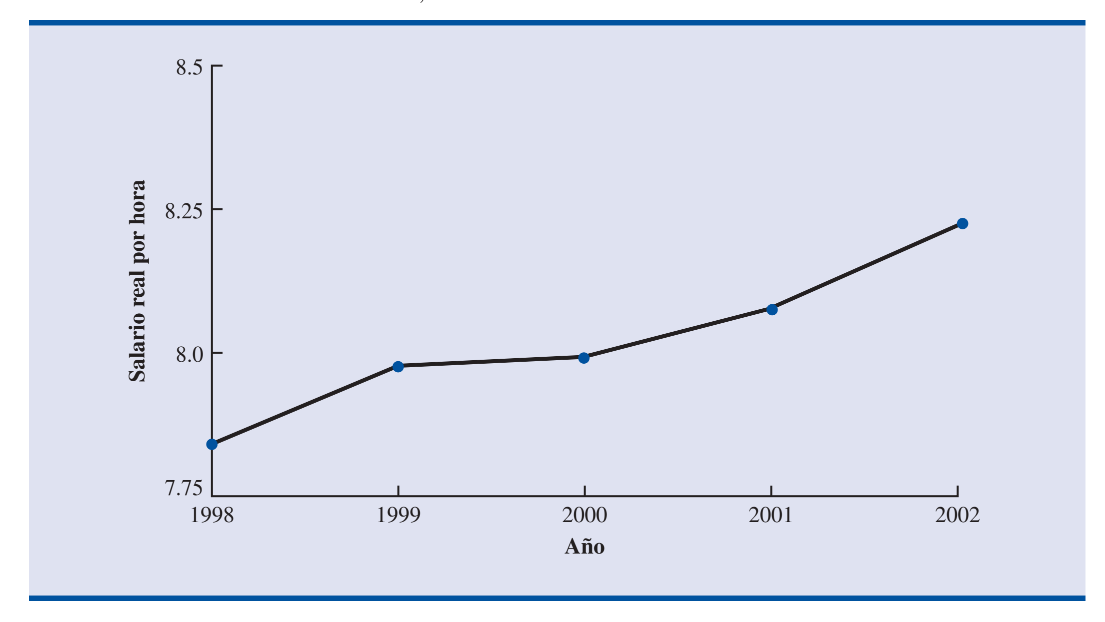
Año	Salario por hora ($)	IPC(base: 1982-1984)
1998	12.78	163.0
1999	13.24	166.6
2000	13.76	172.2
2001	14.31	177.1
2002	14.77	179.9

Fuente: Departamento de Estadística Laboral.

TABLA 17.9 SERIE DEFLACTADA DE LOS SALARIOS POR HORA PARA LOS TRABAJADORES DE LA INDUSTRIA, 1998-2002

Año	Salario por hora deflactado
1998	($12.78/163.0)(100) = $7.84
1999	($13.24/166.6)(100) = $7.95
2000	($13.76/172.2)(100) = $7.99
2001	($14.31/177.1)(100) = $8.08
2002	($14.77/179.9)(100) = $8.21

FIGURA 17.3 SALARIOS REALES POR HORA PARA LOS TRABAJADORES DE LA INDUSTRIA, 1998-2002



cios producidos en un país. Es claro que si el PIB no se deflacta mediante un índice de precios, con el tiempo mostrará aumentos debidos en parte al aumento de los precios. Por tanto, para ajustar el valor total de los bienes y servicios de manera que reflejen los verdaderos cambios en el volumen de bienes y servicios producidos y vendidos, el PIB debe ser calculado con un índice de precios como deflactor. El proceso es similar al visto para calcular los salarios reales.

Ejercicios

Aplicaciones



10. En febrero de 1996, el salario promedio por hora para los trabajadores de la industria era de $11.86; en febrero de 2006 era $16.47. El IPC en febrero de 1996 era 154.9; en 2006 era 198.7.
    a. Deflacte los salarios por hora de 1996 y de 2006 para hallar los salarios reales.
    b. ¿Cuál es el cambio porcentual en los salarios nominales por hora de 1996 a 2006?
    c. ¿Cuál es el cambio porcentual en los salarios reales por hora de 1996 a 2006?

11. Los salarios promedio por hora de los trabajadores de la industria de servicios en los cuatro años desde 2002 hasta 2005 se presentan a continuación. Use los índices de precios al consumidor para deflactar la serie de salarios. Calcule el aumento o la disminución porcentual de los salarios reales y de los salarios nominales, desde 2003 hasta 2005.

Año	Salarios por hora	IPC (base: 1982-1984)
2002	18.52	179.9
2003	18.95	184.0
2004	19.23	188.9
2005	19.46	195.3

Fuente: Departamento de Estadística Laboral.

12. La Oficina de Censos de Estados Unidos publica las siguientes cantidades de pedidos de la industria desde 1999 hasta 2001.

Año	Pedidos de la industria ($ miles de millones)
1999	4032
2000	4218
2001	3971

a. Los índices de precios al consumidor desde 1998 hasta 2002 se presentan en la tabla 17.8. Use esta información para deflactar la serie de pedidos de la industria y dé un comentario respecto al patrón que siguen los pedidos de la industria en términos de dólar constante.
b. Los siguientes índices de precios al consumidor (bienes de consumo terminados) corresponden a los años 1999 a 2001, tome como año base 1982. Use el IPP para deflactar la serie.

Año	IPP (1982 = 100)
1999	133.0
2000	138.0
2001	140.7

c. ¿Qué índice piensa que sea más adecuado como deflactor de los pedidos de la industria, el IPC o el IPP?

13. En la tabla siguiente se presentan volúmenes totales de ventas al menudeo de los Dooley Retail Outlets', para algunos de los años de 1982 a 2002. También se presentan los IPC con base 1982-1984. Deflacte las cifras de los volúmenes de venta con base en el dólar constante de 1982-1984, y haga un comentario sobre el volumen de venta de esta empresa en términos de dólares deflactados.

Año	Ventas al menudeo ($)	IPC (base: 1982-1984)
1982	380 000	96.5
1987	520 000	113.6
1992	700 000	140.3
1997	870 000	160.5
2002	940 000	179.9



Índices de precios: otras consideraciones

En las secciones anteriores se describieron varios métodos para calcular índices de precios, se vio el uso de algunos de los índices más importantes y se presentó un procedimiento para usar los índices de precios en la deflacción de series de tiempo. Para comprender mejor cómo se construyen y se usan los índices de precios, es necesario considerar también otras cuestiones, algunas de las cuales se verán en esta sección.

Selección de los artículos

El propósito principal de un índice de precios es medir la variación, en el transcurso del tiempo, del precio de un determinado conjunto de artículos, productos, etc. Si este conjunto es muy grande, el índice no se puede basar en todos los artículos de tal conjunto; es necesario usar una muestra representativa de estos artículos. Mediante los precios y cantidades de los artículos en la muestra, se espera poder tener una buena idea del comportamiento de los precios de todos los artículos que representa el índice. Por ejemplo, para el índice de precios al consumidor, la cantidad de artículos a considerar, como población de artículos que compra normalmente un consumidor, es de 2000 o más. Pero este índice se basa sólo en las características de precio y cantidad de 400 artículos. La selección de los artículos para calcular este índice no es una tarea trivial. Se requiere estudiar los patrones de compra de los consumidores, así como tener un buen criterio; para seleccionar los 400 artículos no se emplea una muestra aleatoria simple.

Una vez realizada la selección inicial, el grupo de artículos que se usa para calcular el índice se revisa y se modifica periódicamente siempre que haya modificaciones en los patrones de compra. De manera que la decisión de cuáles son los artículos a incluir en un índice es un problema que debe ser resuelto para elaborar el índice y para modificarlo.

Selección de un periodo base

La mayor parte de los índices que se establecen dan el valor 100 al periodo base. Todos los valores futuros del índice son valores en relación con el valor del periodo base. Cuál es el periodo base apropiado para un índice, no es una pregunta que sea fácil de responder. Debe estar basada en el criterio de la persona que elabora el índice.

Muchos de los índices establecidos por el gobierno de Estados Unidos hasta el 2006 utilizan 1982 como periodo base. Como lineamiento general, el periodo base no debe ser un periodo muy alejado del actual. Por ejemplo, un índice de precios al consumidor que tenga como periodo base 1945 sería un índice difícil de entender para la mayoría de los individuos, debido a que la mayoría no está familiarizada con las condiciones de 1945. Así, el periodo base de la mayor parte de los índices se cambia periódicamente por un periodo más reciente. El periodo base para el IPC se cambió en 1988, de 1967 al promedio de 1982-1984. El IPP emplea 1982 como periodo base (es decir, 1982 = 100).

Variaciones en la calidad

El propósito de un índice es medir las variaciones de los precios a lo largo del tiempo. Lo ideal es obtener los datos, en diferentes tiempos, de los precios de los artículos de un mismo conjunto y después calcular el índice. Una suposición básica es que en cada periodo se identifiquen los precios de los mismos artículos. Un problema se presenta cuando la calidad de los productos cambia de un periodo al siguiente. Por ejemplo, un fabricante puede modificar la calidad de un producto de un año a otro, ya sea con materiales menos costosos, al modificar las características del artículo, etc. El precio puede aumentar en los años siguientes, pero es un precio por un producto de menor calidad. En consecuencia, en realidad, el precio puede aumentar más de lo que indica la lista de precios. Es difícil, si no es que imposible, ajustar un índice a la disminución de la calidad de un artículo.

Una mejora importante en la calidad ocasiona un aumento en el precio de un producto. La parte del precio que está relacionada con la mejora de la calidad debe excluirse del cálculo del índice. Sin embargo, ajustar un índice a un aumento de precio que está relacionado con una mejor calidad de un artículo es extremadamente difícil, si no es que imposible.

Al elaborar un índice de precios se acostumbra ignorar cambios pequeños en la calidad de un artículo, los cambios importantes sí deben tomarse en cuenta debido a que éstos pueden modificar, de un periodo a otro, la descripción de un producto. Si la descripción de un producto cambia, el índice debe modificarse para reflejar este cambio; en algunos casos ese producto se elimina del índice.

Sin embargo, en otros casos, una mejora importante de la calidad va acompañada por una disminución del precio. Esta situación, menos común, es lo que ocurrió con las computadoras personales durante los años noventa y los primeros años del siglo XXI.



17.7 Índices de cantidad

Además de los índices de precios descritos en las secciones anteriores, hay otros tipos de índices también útiles. En particular, otra aplicación de los números índice es medir cambios de cantidad a lo largo del tiempo. A este tipo de índices se les conoce como **índices de cantidad**.

Recuerde que en la sección 17.2, al obtener el índice de precios agregados ponderados, para calcular el número índice del periodo t se necesitaron los precios unitarios del periodo base (P_0) y del periodo t (P_t). Mediante la ecuación (17.3) se obtuvo el índice de precios agregados ponderados de la manera siguiente

$$I_t = \frac{\Sigma P_{it} Q_i}{\Sigma P_{i0} Q_i}(100)$$

El numerador, $\Sigma P_{it}Q_i$, representa el valor total de cantidades fijas de los artículos del índice en el periodo t. El denominador, $\Sigma P_{i0}Q_i$, representa el valor total de las mismas cantidades fijas de los artículos del índice en el periodo 0.

El cálculo de un índice de cantidades agregadas ponderadas es similar al del índice de precios agregados ponderados. Las cantidades de cada artículo se miden en el periodo base y en el periodo t; Q_{i0} y Q_{it} representan, respectivamente, estas cantidades del artículo i. Después, estas cantidades se ponderan mediante un precio fijo, el valor agregado o algún otro factor. El "valor agregado" de un producto es el valor de venta menos el costo de las inversiones. La fórmula para calcular el índice de cantidad agregada ponderada para el periodo t es

$$I_t = \frac{\Sigma Q_{it} w_i}{\Sigma Q_{i0} w_i}(100) \tag{17.9}$$

En algunos índices de cantidad, el peso que se usa para el artículo i es el precio en el periodo base (P_{i0}), en cuyo caso, el índice de cantidades agregadas ponderadas es

$$I_t = \frac{\Sigma Q_{it} P_{i0}}{\Sigma Q_{i0} P_{i0}}(100) \tag{17.10}$$

Los índices de cantidad también pueden calcularse con base en cantidades relativas ponderadas. A continuación se presenta una fórmula para esta versión de índices de cantidad

$$I_t = \frac{\sum \frac{Q_{it}}{Q_{i0}}(Q_{i0} P_i)}{\Sigma Q_{i0} P_i}(100) \tag{17.11}$$

Esta fórmula es la versión para cantidades de la fórmula para precios relativos ponderados de la ecuación (17.8), sección 17.3.

El índice de cantidad más conocido es probablemente el **índice de la producción industrial**, elaborado por el consejo de la Reserva Federal de Estados Unidos. Este índice se publica cada mes y el periodo base es al año 2002. Este índice tiene por objeto medir variaciones en el volumen de producción de diversos productos manufacturados, además de productos de minería y de empresas de servicio público. En febrero de 2006 el índice fue 110.9.

Ejercicios

Métodos



14. En la siguiente tabla se presentan datos de las cantidades vendidas de tres artículos en 1995 y en 2006, se dan también los precios de venta de estos artículos en 1995. Calcule el índice de cantidades agregadas ponderadas para 2006.

	Cantidad vendida		
Artículo	**1995**	**2006**	**Precio unitario 1995 ($)**
A	350	300	18.00
B	220	400	4.90
C	730	850	15.00

Aplicaciones



15. Una empresa de transporte de mercancías traslada tres productos de un determinado distribuidor. En la tabla siguiente se presentan los envios de estos tres productos en 1994 y en 2006, así como los precios de 1994.

	Materia prima		Precio/envío
Producto	**1994**	**2006**	**1994**
A	120	95	$1200
B	86	75	$1800
C	35	50	$2000
D	60	70	$1500

Obtenga un índice de cantidades agregadas ponderadas con 1994 como base. Presente un comentario sobre el aumento o la disminución de estas cantidades en el periodo 1994-2006.

16. Un comerciante de automóviles presenta en la siguiente tabla sus ventas de tres modelos en 1992 y en 2006. Calcule las cantidades relativas y úselas para obtener un índice de cantidades agregadas ponderadas para al año 2006, emplee los datos de los dos años.

	Ventas		Precio medio por venta
Modelo	**1992**	**2006**	**(1992)**
Sedan	200	170	$15 200
Sport	100	80	$17 000
Wagon	75	60	$16 800

Resumen

Los índices de precio y la cantidad son medidas importantes de las variaciones de precio y cantidad en la economía y el comercio. Los precios relativos son simplemente el cociente del precio unitario actual de un artículo entre el precio unitario del periodo base, multiplicado por 100; si el valor del precio relativo de un artículo es 100, esto indica que no hay diferencia entre el precio actual y el del periodo base. Los índices de precios agregados son una medida compuesta de la variación general en los precios de un determinado grupo de artículos o productos. Los artículos de un índice de precios agregados suelen ponderarse mediante su cantidad de uso. Un índice de precios agregados ponderados también se calcula al ponderar los precios relativos mediante las cantidades de uso en los artículos del índice.

El índice de precios al consumidor y el índice de precios al productor son dos índices muy citados y sus años base son, respectivamente, 1982-1984 y 1982. El promedio industrial Dow Jones es otro índice de precios muy citado. Este índice es una suma ponderada de los precios de 30 acciones ordinarias de empresas grandes. A diferencia de muchos otros índices, este índice no se da como porcentaje del valor de algún año base.

Con frecuencia los índices de precios se usan para deflactar algunas otras series económicas que se miden a lo largo del tiempo. Se vio el uso del IPC para deflactar los salarios por hora y obtener un índice de salarios reales. Consideraciones importantes para obtener un número índice son la selección de los artículos a incluir en el índice, la selección del periodo base para el índice y el ajuste a las variaciones en la cantidad. También se revisaron brevemente los índices de cantidad y se mencionó el índice de la producción industrial como un importante índice de cantidad.

Glosario

Precio relativo Índice de precio para un determinado artículo que se calcula al dividir un precio unitario actual entre un precio unitario del periodo base y multiplicar el resultado por 100.
Índice de precios agregados Índice de precio compuesto que se basa en los precios de un grupo de artículos.
Índice de precios agregados ponderados Índice en el cual los precios de los artículos que lo componen son ponderados mediante su importancia relativa.
Índice de Laspeyres Índice de precios agregados ponderados en el que el peso de cada artículo es su cantidad en el periodo base.
Índice de Paasche Índice de precios agregados ponderados en el que el peso de cada artículo es su cantidad en el periodo actual.
Índice de precios al consumidor (IPC) Índice de precios mensual que usa las variaciones de precio en la canasta de mercado de los bienes y servicios de consumo para medir las variaciones en los precios al consumidor a lo largo del tiempo.
Índice de precios al productor (IPP) Índice de precios mensual diseñado para medir variaciones en los precios de bienes vendidos en mercados primarios (esto es, la primera compra de una materia prima en un mercado no minorista).
Promedios Dow Jones Índice de precios agregados diseñado para mostrar tendencias de precios y movimientos en la Bolsa de Cambio de Nueva York.
Índice de cantidad Índice que tiene por objeto medir cambios de cantidades a lo largo del tiempo.
Índice de la producción industrial Un índice de cantidad diseñado para medir variaciones, a lo largo del tiempo, en el volumen físico o en los niveles de producción de los bienes industriales.

Fórmulas clave

Precio relativo del periodo *t*

$$\frac{\text{Precio del periodo } t}{\text{Precio del periodo base}}(100) \tag{17.1}$$

Índice de precios agregados no ponderados del periodo *t*

$$I_t = \frac{\Sigma P_{it}}{\Sigma P_{i0}}(100) \tag{17.2}$$

Índice de precios agregados ponderados del periodo *t*

$$I_t = \frac{\Sigma P_{it} Q_i}{\Sigma P_{i0} Q_i}(100) \tag{17.3}$$

Promedio ponderado de precios relativos

$$I_t = \frac{\sum \frac{P_{it}}{P_{i0}}(100) w_i}{\Sigma w_i} \tag{17.6}$$

Factor de ponderación en la ecuación (17.6)

$$w_i = P_{i0} Q_i \tag{17.7}$$

Índice de cantidades agregadas ponderadas

$$I_t = \frac{\Sigma Q_{it} w_i}{\Sigma Q_{i0} w_i}(100) \tag{17.9}$$

Ejercicios complementarios

17. Los precios medianos de casas habitación nuevas de 1998 al 2001 son los siguientes (*Statistical Abstract of the United States,* 2002).

Año	Precio ($ miles)
1998	152.5
1999	161.0
2000	169.0
2001	175.2

a. Con 1998 como año base, obtenga un índice de precios, correspondiente a estos años, para casas habitación nuevas.
b. Con 1999 como año base, obtenga un índice precios de casas habitación nuevas para estos años.

18. Los siguientes son datos de la empresa Nickerson Manufacturing sobre las cantidades de sus pedidos y el costo unitario de cada uno de sus productos:

Productos	Cantidades en el periodo base (2003)	Costo unitario medio del pedido ($) 2003	2006
A	2000	10.50	15.90
B	5000	16.25	32.00
C	6500	12.20	17.40
D	2500	20.00	35.50

a. Calcule el precio relativo de cada producto.
b. Calcule un índice de precios agregados ponderados que refleje la variación del costo de los pedidos en estos cuatro años.

19. Utilice los datos del ejercicio 18 para calcular el índice Paasche de los costos de los pedidos si en el 2006 las cantidades en los pedidos son 4000, 3000, 7500 y 3000 para cada uno de los productos.

20. Boran Stockbrokers, Inc., selecciona cuatro acciones con el propósito de obtener su propio índice para el comportamiento del mercado de acciones. A continuación se dan los precios por acción en el 2004, que es el periodo base, en enero de 2006 y en marzo de 2006. Las cantidades del año base se fijan de acuerdo con los volúmenes históricos de las cuatro acciones.

			Precio por acción ($)		
Acción	**Industria**	**Cantidad en 2004**	**Base: 2004**	**Enero 2006**	**Marzo 2006**
A	Petrolera	100	31.50	22.75	22.50
B	De la computación	150	65.00	49.00	47.50
C	Del acero	75	40.00	32.00	29.50
D	Inmobiliaria	50	18.00	6.50	3.75

Use el periodo base, 2004, para calcular el índice Boran correspondiente a enero de 2006 y a marzo de 2006. Haga un comentario sobre lo que dice este índice acerca de lo que ocurre en el mercado de acciones.

21. Calcule los precios relativos de las cuatro acciones que en el ejercicio 20 se emplean para el índice Boran. Use los agregados ponderados de los precios relativos para calcular los índices Boran correspondientes a enero de 2006 y a marzo de 2006.

22. Considere la información siguiente sobre precios relativos y cantidades referentes a la producción de grano en Iowa (*Statistical Abstract of the United States,* 2002).

Producto	**Cantidades en 1991 (millones de bushels)**	**Precio base por bushel ($)**	**Precios relativos 1991-2001**
Maíz	1427	2.30	91
Frijol de soya	350	5.51	78

¿Cuál es el índice de precios agregados ponderados del 2001 para los granos de Iowa?

23. A continuación se presentan datos sobre el consumo de las frutas y los precios en 1988 y en 2001.

Fruta	**Consumo per cápita, 1998 (libras)**	**Precio 1988 ($/libras)**	**Precio 2001 ($/libras)**
Plátanos	24.3	0.41	0.51
Manzanas	19.9	0.71	0.87
Naranjas	13.9	0.56	0.71
Peras	3.2	0.64	0.98

a. Calcule el precio relativo de cada producto.
b. Calcule el índice de precios agregados ponderados de estos productos. Haga un comentario sobre la variación de los precios de las frutas en este lapso de 13 años.

24. A continuación se presentan los salarios iniciales de los asistentes de profesor de administración en una universidad. Use los IPC para deflactar los salarios a dólar constante. Haga un comentario sobre la tendencia de los salarios en la educación de acuerdo con lo que indican estos datos.

Año	**Salario inicial ($)**	**IPC (base: 1982-1984)**
1970	14 000	38.8
1975	17 500	53.8
1980	23 000	82.4
1985	37 000	107.6
1990	53 000	130.7
1995	65 000	152.4
2000	80 000	172.2
2005	110 000	195.3

25. Para una determinada acción se dan a continuación los precios por acción en cinco años consecutivos, así como los IPC que tienen como periodo base 1982-1984.

Año	**Precio por acción ($)**	**IPC (base: 1982-1984)**
2001	51.00	177.1
2002	54.00	179.9
2003	58.00	184.0
2004	59.50	188.9
2005	59.00	195.3

Deflacte la serie de precios de estas acciones y haga un comentario sobre la inversión en estas acciones.

26. En la tabla siguiente se presentan cantidad y valor de los productos de una empresa fabricante en los años 2002 y 2006. Con estos datos calcule un índice de cantidades agregadas ponderadas. Dé un comentario sobre el significado de este índice de cantidad.

	Cantidad		
Producto	**2002**	**2006**	**Valores ($)**
A	800	1200	30.00
B	600	500	20.00
C	200	500	25.00

CAPÍTULO 18



Pronóstico

CONTENIDO

LA ESTADÍSTICA *en* LA PRÁCTICA

OCCUPATIONAL HEALTH CLINIC DE NEVADA*
SPARKS, NEVADA

La Occupational Health Clinic de Nevada es una clínica privada que se encuentra en Sparks, Nevada. La clínica se ha especializado en medicina industrial y opera en el mismo sitio desde hace más de 20 años. La clínica había tenido una etapa de crecimiento rápido en la que su facturación mensual creció de $57 000 a más de $300 000, durante 26 meses, cuando un incendio consumió el edificio principal de la clínica.

La póliza de seguro de la clínica cubría los daños materiales y al equipo, así como los ingresos durante la interrupción de su funcionamiento normal. Establecer la declaración de daños materiales fue relativamente sencillo, ya que consistió en determinar el valor de los mismos y de la pérdida de equipo a causa del fuego. Sin embargo, determinar el valor del ingreso durante los siete meses que se necesitaron para reconstruir la clínica fue un asunto complicado que requirió de negociaciones entre los dueños de la clínica y la empresa aseguradora. No había reglas preestablecidas para calcular "qué hubiera pasado" con la facturación de la clínica de no haber ocurrido el incendio.

Para estimar los ingresos perdidos, la clínica empleó un método de pronóstico para proyectar el crecimiento que hubieran tenido los ingresos durante los siete meses que la clínica permaneció cerrada. La historia de la facturación antes del incendio suministró la base para obtener un modelo de pronóstico de tendencia lineal y componentes estacionales como los que se discuten en este capítulo. El modelo de pronóstico permitió a la clínica establecer una estimación exacta de la pérdida, la cual fue finalmente aceptada por la empresa de seguros.



Un incendio hizo que la Occupational Health Clinic de Nevada se cerrara durante siete meses. © PhotoDisc/Getty Images.

*Los autores agradecen a Bard Betz, director de operaciones y a Curtis Brauer, asistente ejecutivo de la Occupational Health Clinic de Nevada por proporcionar este artículo para *La estadística en la práctica*.

Un aspecto esencial en la administración de cualquier organización es la planeación para el futuro. En efecto, el éxito a largo plazo de una organización está estrechamente relacionado con la capacidad que tenga la administración de anticipar el futuro y elaborar estrategias adecuadas. Criterio, intuición y atención al entorno económico permiten que un administrador tenga una idea aproximada de lo que puede ocurrir en el futuro. Sin embargo, no resulta fácil convertir una idea aproximada en un número que pueda representar el volumen de ventas del trimestre próximo o el costo de la materia prima en el año siguiente. El objetivo de este capítulo es presentar varios métodos para obtener pronósticos.

La mayor parte de las empresas pueden pronosticar la demanda total de sus productos con un error menor a 5%. Sin embargo, al pronosticar la demanda de uno solo de los productos es posible tener errores significativamente mayores.

Un pronóstico es simplemente un pronóstico de lo que ocurrirá en el futuro. Los administradores deben aceptar que, sea cual sea la técnica que usen, no podrán obtener pronósticos perfectos.

Suponga que se le pide dar un pronóstico trimestral del volumen de ventas durante el año próximo de un determinado producto. El pronóstico trimestral que dé afectará el programa de producción, la compra de materias primas, las políticas de inventario y el monto de las ventas. En consecuencia, un mal pronóstico puede llevar a una mala planeación, lo cual incrementará los costos de la empresa. ¿Cómo proceder para dar un pronóstico trimestral del volumen de ventas?

Seguramente habrá que revisar los datos de las ventas del producto en los periodos pasados. Utilizar estos datos históricos permite identificar el volumen general de ventas y si, con el tiempo, existe alguna tendencia de aumento o disminución en el volumen de ventas. Mediante una revisión más cuidadosa de los datos podría determinarse, por ejemplo, si las ventas siguen un patrón estacional que se manifieste en un aumento de las ventas en el tercer trimestre de cada año y volúmenes de ventas que tocan fondo en el primer trimestre. Revisar los datos históricos facilita entender mejor el patrón de las ventas en el pasado, lo que lleva a mejores pronósticos de las ventas futuras del producto.

Los datos históricos forman una serie de tiempo. Una **serie de tiempo** es un conjunto de observaciones de una variable medida en puntos sucesivos en el tiempo o en periodos de tiempo sucesivos. En este capítulo se presentan varios métodos para el análisis de series de tiempo. El objetivo de tal análisis es obtener un buen pronóstico o predicción de los valores futuros de una serie de tiempo.

Los métodos de pronóstico se clasifican como cualitativos y cuantitativos. Los métodos cuantitativos de pronóstico se suelen usar cuando 1) se cuenta con información del pasado acerca de la variable que se desea pronosticar, 2) esa información se puede cuantificar y 3) es razonable pensar que el patrón seguido en el pasado continuará en el futuro. En tales casos es posible obtener un buen pronóstico mediante un método de series de tiempo o un método causal.

Si los datos históricos están restringidos a valores pasados de la variable, al método de pronóstico se le conoce como un *método de series de tiempo*. El objetivo de un método de series de tiempo es descubrir en los datos históricos un patrón para después extrapolar ese patrón al futuro; el pronóstico se basa únicamente en los valores de la variable en el pasado o en errores de pronóstico en el pasado. En este capítulo se estudian tres métodos de series de tiempo: suavizamiento (promedios móviles, promedios movibles ponderados y suavizamiento exponencial), proyección de tendencia y proyección de tendencia ajustada a la influencia estacional.

Los métodos de pronóstico causal están sustentados en la suposición de que la variable a pronosticar tiene una relación de causa y efecto con otra u otras variables. En este capítulo se verá el uso del análisis de regresión como método de pronóstico causal. Por ejemplo, los gastos en publicidad suelen influir en los volúmenes de ventas de muchos artículos, de manera que se puede emplear el análisis de regresión para obtener una ecuación que muestre cómo es la relación entre estas dos variables. Así, una vez que se establezca la cantidad presupuestada para la publicidad del periodo siguiente, se podrá sustituir este valor en la ecuación y obtener una predicción o pronóstico para el volumen de ventas de ese periodo. Observe que si se utiliza un método de series de tiempo para obtener un pronóstico, los gastos en publicidad no son tomados en consideración; es decir, cuando se emplea un método de series de tiempo el pronóstico se sustenta únicamente en las ventas del pasado.

En los métodos cualitativos, para obtener un pronóstico, suele necesitarse el criterio de un experto. Por ejemplo, un panel de expertos puede elaborar un pronóstico consensual para el tipo de interés preferencial que estará vigente durante un año a partir de ahora. Los métodos cualitativos presentan ventajas cuando la información acerca de la variable que se pronostica no puede cuantificarse y cuando no se cuenta con datos históricos o cuando los datos históricos con los que se cuenta no son aplicables. En la figura 18.1 se presenta una visión general de los métodos de pronóstico.



18.1 Componentes de una serie de tiempo

El patrón o el comportamiento que siguen los datos de una serie de tiempo se debe a diversos componentes. Por lo general se supone que son cuatro los componentes que se combinan para dar los valores de una serie de tiempo: de tendencia, cíclico, estacional e irregular. A continuación se verá cada uno de estos cuatro componentes.

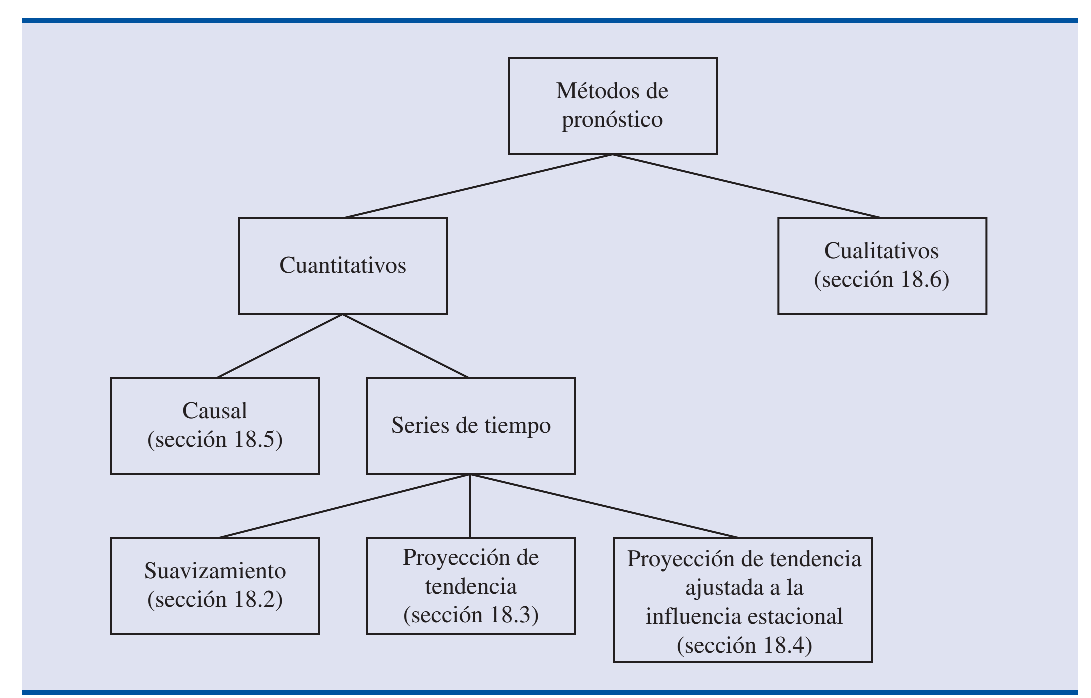
Componente de tendencia

En el análisis de las series de tiempo, las mediciones pueden hacerse cada hora, diario, a la semana, cada mes, anualmente o en cualquier otro intervalo regular de tiempo.* Aunque los datos de las series de tiempo suelen mostrar fluctuaciones aleatorias, las series de tiempo también muestran un desplazamiento o movimiento gradual hacia valores relativamente altos o bajos a través de un lapso largo. A este desplazamiento gradual de la serie de tiempo se le conoce como la **tendencia** de la serie de tiempo; este desplazamiento o tendencia suele deberse a factores de largo plazo como variaciones en las características demográficas de la población, en la tecnología o en las preferencias del público.

Por ejemplo, un fabricante de cámaras fotográficas encuentra que cada mes existe una variación sustancial en la cantidad de cámaras que vende. Sin embargo, al observar las ventas de los

*Aquí el estudio se limitará a las series de tiempo con valores tomados a intervalos regulares. Los casos en que las observaciones no se hacen a intervalos regulares quedan fuera del alcance de este libro.

FIGURA 18.1 VISIÓN GENERAL DE LOS MÉTODOS DE PRONÓSTICO



últimos 10 a 15 años, este fabricante se percata de que ha habido un aumento gradual del volumen de ventas. Suponga que en 1996 el volumen de ventas fue aproximadamente de 17 000 cámaras; en 2001, 23 000 cámaras y en el 2006, 25 000 cámaras. Este crecimiento gradual de las ventas a lo largo del tiempo representa una tendencia ascendente de esta serie de tiempo. En la figura 18.2 se muestra una línea recta que puede ser una buena aproximación a la tendencia observada en las ventas de las cámaras fotográficas. Aunque la tendencia en las ventas de las cámaras parece ser lineal y creciente, en una serie de tiempo pueden presentarse tendencias que se describen mejor mediante algún otro patrón.

FIGURA 18.2 TENDENCIA LINEAL EN LAS VENTAS DE CÁMARAS FOTOGRÁFICAS

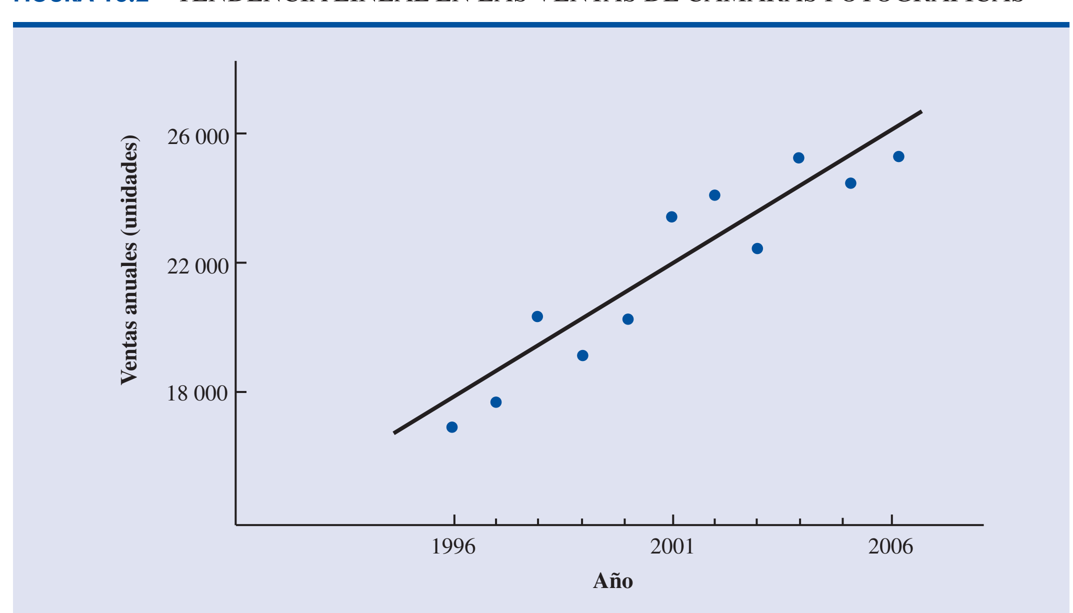


FIGURA 18.3 EJEMPLOS DE ALGUNOS PATRONES POSIBLES EN UNA SERIE DE TIEMPO

Volumen

Tiempo

A. No lineal

Volumen

Tiempo

B. Tendencia lineal decreciente

Volumen

Tiempo

C. Sin tendencia

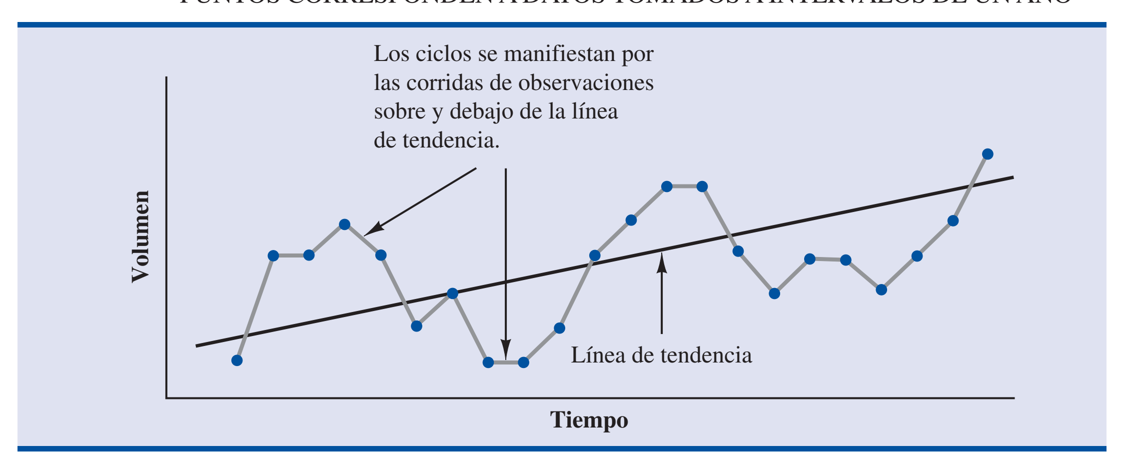
En la figura 18.3 se muestran otros patrones que puede seguir la tendencia de una serie de tiempo. En la gráfica A se muestra una tendencia no lineal; en este caso, al inicio se observa poco crecimiento en la serie de tiempo y al final una estabilización. Esta tendencia puede ser una buena aproximación a las ventas de un producto desde su introducción, seguida por una etapa de crecimiento y llegando al final a un periodo de saturación del mercado. La tendencia de disminución lineal que se observa en la gráfica B se encuentra en series de tiempo que observan una disminución creciente a lo largo del tiempo. La línea horizontal que se observa en la gráfica C representa una serie de tiempo en la que no hay un aumento ni una disminución consistentes a lo largo del tiempo y, por tanto, en la que no hay tendencia alguna.

Componente cíclico

Aunque una serie de tiempo puede tener una tendencia a través de lapsos largos, no todos los valores futuros de la serie de tiempo caerán exactamente sobre la línea de tendencia. Las series de tiempo suelen mostrar secuencias de puntos que caen de manera alternante arriba y abajo de la línea de tendencia. Toda sucesión recurrente de puntos que caiga abajo y arriba de la línea de tendencia y que dure más de un año puede atribuirse al **componente cíclico** de la serie de tiempo. En la figura 18.4 se muestra la gráfica de una serie de tiempo con un componente cíclico obvio. Estas observaciones fueron tomadas en intervalos de un año.

Muchas series de tiempo muestran un comportamiento cíclico, con observaciones que caen de manera regular abajo y arriba de la línea de tendencia. Por lo general, este componente de las series de tiempo es debido a movimientos cíclicos multianuales de la economía. Por ejemplo, periodos de inflación moderada seguidos de periodos de inflación rápida pueden hacer que la serie

FIGURA 18.4 COMPONENTES DE TENDENCIA Y CÍCLICOS DE UNA SERIE DE TIEMPO. LOS PUNTOS CORRESPONDEN A DATOS TOMADOS A INTERVALOS DE UN AÑO



de tiempo alterne hacia arriba y hacia abajo de la línea general de tendencia creciente (por ejemplo, una serie de tiempo sobre el costo de la vivienda).

Componente estacional

Mientras los componentes cíclico y de tendencia de las series de tiempo se identifican tras el análisis de las variaciones multianuales en los datos históricos, en muchas series de tiempo se observa un patrón permanente en lapsos de un año. Por ejemplo, un fabricante de albercas espera tener pocas ventas durante los meses de otoño e invierno y sus mayores ventas en los meses de primavera y verano. Los fabricantes de equipo para remover la nieve y los fabricantes de ropa de invierno esperan exactamente lo contrario. Por tanto, es natural que al componente de las series de tiempo que representan la variabilidad en los datos debida a la influencia estacional se le conozca como **componente estacional**. Aunque por lo general se considera que las variaciones estacionales son variaciones que se presentan durante el lapso de un año, el componente estacional también se usa para representar cualquier variación que se presente con regularidad en un lapso menor que un año. Por ejemplo, en el volumen de tráfico diario, en el lapso de un día, se observa una conducta "estacional", en donde los valores máximos se presentan en las horas pico, y durante el resto del día y al comienzo de la noche un flujo moderado, y un flujo ligero desde la media noche hasta las primeras horas de la mañana.

Componente irregular

El **componente irregular** de una serie de tiempo es el factor residual o el factor que da cuenta de las desviaciones de los valores reales de la serie de tiempo de los valores que se esperan al considerar los efectos de los componentes de tendencia, cíclicos y estacionales. Este componente irregular es ocasionado por factores a corto plazo, imprevistos y no recurrentes que afectan a la serie de tiempo. Dado que este componente da cuenta de la variabilidad aleatoria en una serie de tiempo, es un componente impredecible. No es posible predecir su efecto sobre la serie de tiempo.



Métodos de suavizamiento

En el ambiente de la manufactura suelen requerirse pronósticos mensuales o semanales para miles de artículos. Por tanto, al elegir una técnica de suavizamiento se requiere sencillez y facilidad de uso. Para las técnicas que se presentan en esta sección los requerimientos de datos son mínimos y las técnicas son fáciles de usar y de entender.

En esta sección se estudian tres métodos de pronóstico: promedios móviles, promedios móviles ponderados y suavizamiento exponencial. Estos métodos tienen por objeto suavizar las fluctuaciones aleatorias ocasionadas por el componente irregular de la serie de tiempo, razón por la que se les conoce como métodos de suavizamiento. Los métodos de suavizamiento son adecuados para series de tiempo estables; es decir, para aquellas series que no muestran efectos importantes de tendencia, cíclicos o estacionales porque se adaptan muy bien a los cambios en el nivel de la serie de tiempo. Sin embargo, sin alguna modificación, no funcionan muy bien cuando hay variaciones importantes de tendencia, cíclicas o estacionales.

Los métodos de suavizamiento son fáciles de utilizar y, por lo general, se obtiene una buena exactitud en pronósticos a corto plazo, como, por ejemplo, pronósticos para el periodo siguiente. Uno de estos métodos, el suavizamiento exponencial, tiene requerimientos mínimos de datos por lo que es un método adecuado cuando se requiere de pronósticos para una gran número de artículos.

Promedios móviles

En el método de los **promedios móviles**, para pronosticar el periodo siguiente, se emplea el promedio de los valores de los n datos más recientes de la serie de tiempo. El cálculo de un promedio móvil se hace como sigue.

PROMEDIO MÓVIL

$$\text{Promedio móvil} = \frac{\Sigma(\text{de los } n \text{ datos más recientes})}{n} \tag{18.1}$$

TABLA 18.1

SERIE DE TIEMPO DE LAS VENTAS DE GASOLINA

Semana	Ventas (miles de galones)
1	17
2	21
3	19
4	23
5	18
6	16
7	20
8	18
9	22
10	20
11	15
12	22



Gasoline

El término *móvil* se usa porque cada vez que en la serie de tiempo hay una nueva observación, ésta sustituye a la observación más antigua que se emplee en la ecuación (18.1) y se calcula un nuevo promedio. De esta manera, el promedio se modifica, o se mueve, cada vez que se tiene una nueva observación.

Para ilustrar el método de los promedios móviles, considere los datos de 12 semanas que se presentan en la tabla 18.1 y en la figura 18.5. En esa tabla se presenta el número de galones de gasolina vendidos en las últimas 12 semanas en una gasolinera de Bennington, Vermont. En la figura 18.5 se observa que aunque existe una variabilidad aleatoria, la serie de tiempo parece ser estable a lo largo del tiempo; por lo que se pueden emplear los métodos de suavizamiento de esta sección.

Para emplear este método en el pronóstico de las ventas de gasolina, primero hay que decidir cuántos valores se usarán para calcular los promedios móviles. Aquí, por ejemplo, se calcularán promedios móviles de tres semanas. El promedio móvil de las ventas de gasolina correspondiente a las tres primeras semanas es

$$\text{Promedio móvil (semanas 1-3)} = \frac{17 + 21 + 19}{3} = 19$$

Este promedio móvil se usa como pronóstico para la semana 4. Como el verdadero valor observado en la semana 4 es 23, el error de pronostico es $23 - 19 = 4$. En un pronóstico el error es la diferencia entre el valor observado en la serie de tiempo y el valor obtenido como pronóstico.

El siguiente promedio móvil de tres semanas es

$$\text{Promedio móvil (semanas 2-4)} = \frac{21 + 19 + 23}{3} = 21$$

FIGURA 18.5 SERIE DE TIEMPO DE LAS VENTAS DE GASOLINA

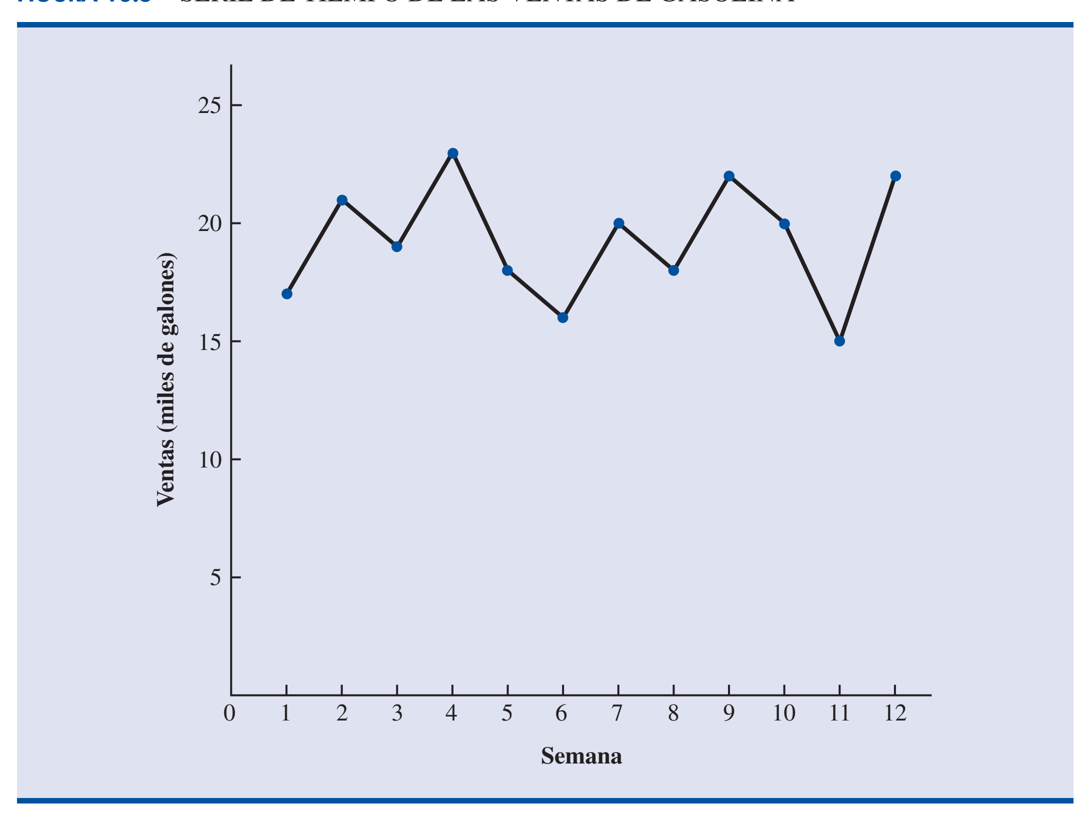


TABLA 18.2 RESUMEN DE LOS CÁLCULOS DE LOS PROMEDIOS MÓVILES DE TRES SEMANAS

Semana	Valores de la serie de tiempo	Pronóstico con el promedio móvil	Error de pronóstico	Error de pronóstico al cuadrado
1	17			
2	21			
3	19			
4	23	19	4	16
5	18	21	−3	9
6	16	20	−4	16
7	20	19	1	1
8	18	18	0	0
9	22	18	4	16
10	20	20	0	0
11	15	20	−5	25
12	22	19	3	9
		Totales	0	92

Por tanto, el pronóstico para la semana 5 es 21. El error correspondiente a este pronóstico es $18 - 21 = -3$. El error de pronóstico puede ser positivo o negativo según si el pronóstico es muy bajo o muy alto. En la tabla 18.2 y en la figura 18.6 se presenta un resumen de los promedios móviles con tres semanas para la serie de tiempo de las ventas de gasolina.

La exactitud del pronóstico no es la única consideración importante. Algunas veces, en métodos más exactos, se requieren datos sobre series de tiempo relacionadas, datos que son costosos o difíciles de obtener. Por lo general, en un pronóstico se tiene que sacrificar costo o exactitud.

Exactitud del pronóstico Al elegir el método de pronóstico es importante considerar la exactitud del método. Es claro que se desea que el error de pronóstico sea pequeño. En las dos últimas columnas de la tabla 18.2 se encuentran los errores de pronóstico y los cuadrados de los errores de pronóstico, los cuales sirven para obtener una medida de la exactitud del pronóstico. Respecto de la serie de tiempo de las ventas de gasolina, se usa la última columna de la tabla 18.2 para calcular el promedio de la suma de los cuadrados de los errores. Así se obtiene

$$\text{Promedio de la suma de los errores al cuadrado} = \frac{92}{9} = 10.22$$

Al promedio de la suma de los errores al cuadrado se le conoce como **cuadrado medio debido al error (CME)**. Este CME suele usarse como medida de la exactitud del método de pronóstico y es la medida que se usará en este capítulo.

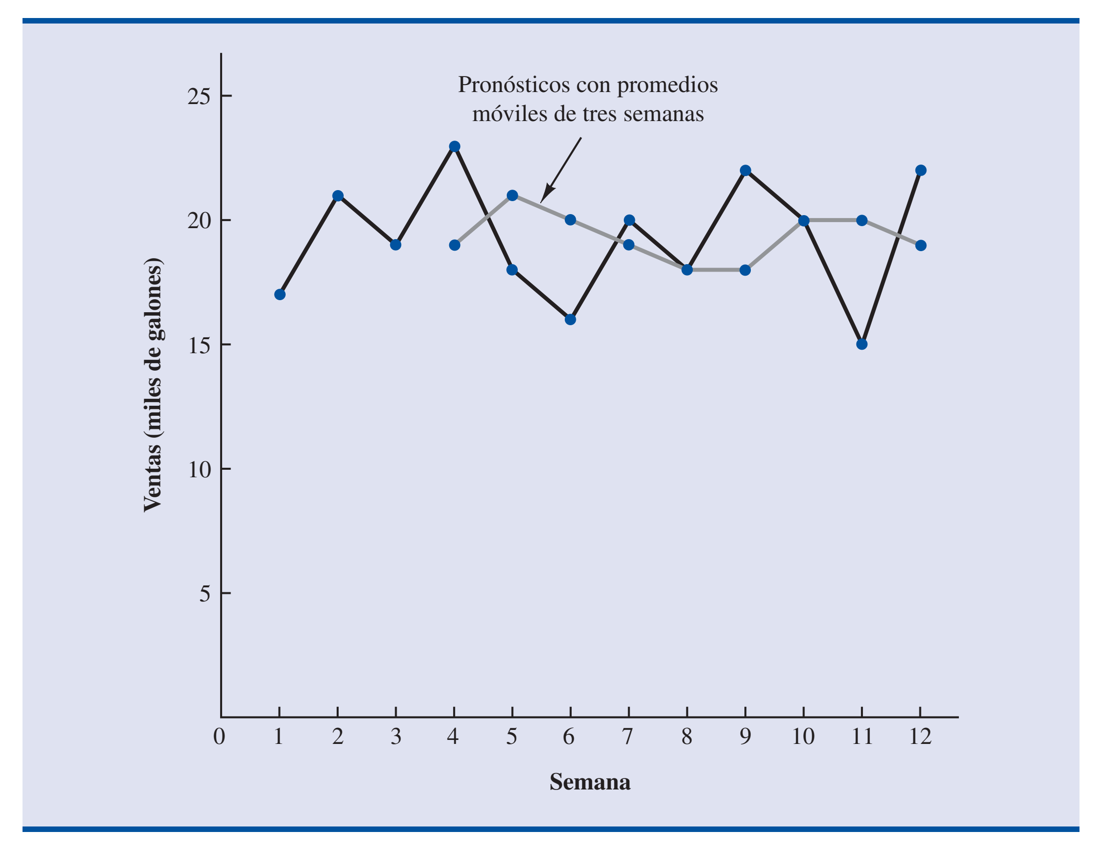
Al final de esta sección, en el ejercicio 2, se pide calcular promedios móviles de cuatro y cinco semanas para las ventas de gasolina.

Como ya se dijo, para usar el método de los promedios móviles, hay que decidir, primero, cuántos datos se emplearán para calcular los promedios móviles. Es claro que para una determinada serie de tiempo, con promedios móviles de longitudes diferentes se obtendrán diferentes exactitudes en el pronóstico de la serie de tiempo. Una manera de decidir cuántos datos emplear en el cálculo de los promedios móviles es determinar, mediante el método de prueba y error, la longitud con la que se minimiza el CME. Si se está dispuesto a asumir que la longitud que es la mejor para el pasado será también la mejor para el futuro, el siguiente valor de la serie de tiempo se pronosticará mediante la cantidad de datos que minimicen el CME en los datos históricos.

Promedios móviles ponderados

En el método de los promedios móviles, a todos los datos que se emplean en el cálculo se les da el mismo peso. En una variación conocida como **promedios móviles ponderados**, a cada uno de los valores de los datos se le da un peso diferente y, después, se calcula el promedio ponderado

FIGURA 18.6 SERIE DE TIEMPO DE LAS VENTAS DE GASOLINA Y PRONÓSTICOS CON PROMEDIOS MÓVILES DE TRES SEMANAS



de los valores de los n datos más recientes para obtener el pronóstico. En la mayoría de las veces, a la observación más reciente se le da el mayor peso y los pesos disminuyen conforme los datos son más antiguos. Por ejemplo, con los datos de las ventas de gasolina se calcula un promedio móvil ponderado de tres semanas: a la observación más reciente se le da un peso que sea el triple del que se le dé a la observación más antigua y a la observación intermedia un peso que sea el doble del de la observación más antigua. Como promedio de la semana 4 se tiene:

$$\text{Pronóstico para la semana 4} = \tfrac{1}{6}(17) + \tfrac{2}{6}(21) + \tfrac{3}{6}(19) = 19.33$$

Observe que en un promedio móvil ponderado, la suma de los pesos es igual a 1. En realidad, la suma de los pesos en el promedio móvil simple también fue igual a 1: cada peso fue de 1/3. Sin embargo, recuerde que con el promedio móvil simple o no ponderado el pronóstico fue 19.

Exactitud del pronóstico Para usar el método de promedios móviles ponderados primero se debe establecer el número de datos a usar para calcular los promedios móviles ponderados y después elegir los pesos para cada uno de los datos. En general, si se cree que el pasado reciente sea un mejor predictor del futuro que el pasado distante, habrá que dar pesos mayores a las observaciones más recientes. Sin embargo, si la serie de tiempo es muy variable, puede ser mejor elegir pesos aproximadamente iguales para todos los datos. Note que el único requerimiento es que la suma de los pesos sea igual a 1. Para estimar si con una determinada combinación de número de datos y pesos, se obtiene un pronóstico más exacto que con otra combinación, se seguirá usando el criterio de CME como medida de la exactitud del pronóstico. Es decir, si se supone que la combinación que es mejor para el pasado también será la mejor para el futuro, para pronosticar el valor siguiente de la serie de tiempo se empleará la combinación de número de datos y pesos, que minimice el CME de la serie de tiempo histórica.

Suavizamiento exponencial

El suavizamiento exponencial es sencillo y tiene pocos requerimientos de datos, por lo que es un método no costoso para las empresas que, en cada periodo, tienen que hacer una gran cantidad de predicciones.

En el **suavizamiento exponencial** se usa un promedio ponderado de los valores pasados de la serie de tiempo; es un caso especial del método de promedios ponderados móviles; en este caso sólo hay que elegir un peso, el peso para la observación más reciente. Los pesos para los demás datos se calculan automáticamente y son más pequeños a medida que los datos son más antiguos. A continuación se presenta el modelo de suavizamiento exponencial básico.

MODELO DE SUAVIZAMIENTO EXPONENCIAL

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t \quad \textbf{(18.2)}$$

donde

F_{t+1} = pronóstico para el periodo $t + 1$ de la serie de tiempo
Y_t = valor real en el periodo t de la serie de tiempo
F_t = pronóstico para el periodo t de la serie de tiempo
α = constante de suavizamiento $(0 \leq \alpha \leq 1)$

La ecuación (18.2) muestra que el pronóstico para el periodo $t + 1$ es un promedio ponderado del valor real en el periodo t y del valor pronosticado para el periodo t; observe, en particular, que el peso dado al valor real del periodo t es α y el peso dado al valor pronosticado para el periodo t es $1 - \alpha$. En una serie de tiempo, como ejemplo, de tres datos: Y_1, Y_2 y Y_3 se demostrará que el pronóstico obtenido mediante suavizamiento exponencial para cualquier periodo es un promedio ponderado de *todos los valores reales anteriores* de la serie de tiempo. Para empezar, sea F_1 igual al valor real de la serie de tiempo en el periodo 1; es decir $F_1 = Y_1$. Por tanto, el pronóstico para el periodo 2 es

$$\begin{aligned} F_2 &= \alpha Y_1 + (1 - \alpha)F_1 \\ &= \alpha Y_1 + (1 - \alpha)Y_1 \\ &= Y_1 \end{aligned}$$

De tal manera que el pronóstico obtenido mediante suavizamiento exponencial para el periodo 2 es igual al valor real de la serie de tiempo para el periodo 1.

El pronóstico para el periodo 3 es

$$F_3 = \alpha Y_2 + (1 - \alpha)F_2 = \alpha Y_2 + (1 - \alpha)Y_1$$

En último lugar, al sustituir la expresión para F_3 en la expresión para F_4 se obtiene

$$\begin{aligned} F_4 &= \alpha Y_3 + (1 - \alpha)F_3 \\ &= \alpha Y_3 + (1 - \alpha)[\alpha Y_2 + (1 - \alpha)Y_1] \\ &= \alpha Y_3 + \alpha(1 - \alpha)Y_2 + (1 - \alpha)^2 Y_1 \end{aligned}$$

El término suavizamiento exponencial se debe al carácter exponencial de los pesos que se emplean para los datos históricos.

Por tanto, F_4 es un promedio ponderado de los tres primeros valores de la serie de tiempo. La suma de los coeficientes o pesos de Y_1, Y_2 y Y_3 es igual a uno. Mediante un argumento similar se puede demostrar que, en general, cualquier pronóstico F_{t+1} es un promedio ponderado de los valores previos de la serie de tiempo.

A pesar de que con el suavizamiento exponencial se obtiene un pronóstico que es el promedio ponderado de todas las observaciones pasadas, no es necesario conservar todos los datos pasados para calcular el pronóstico para el periodo siguiente. En efecto, una vez elegida la **constante de suavizamiento** α, sólo se necesitan dos informaciones para calcular el pronóstico. En la ecuación (18.2) se observa que dada una α, para calcular el pronóstico en el periodo $t + 1$ sólo se necesita conocer el valor real y el valor pronosticado de la serie de tiempo para el periodo t, es decir, Y_t y F_t.

Para ilustrar el uso del método de suavizamiento exponencial para obtener pronósticos, considere la serie de tiempo de los precios de la gasolina presentada en la tabla 18.1 y en la figura 18.5. Como ya se mostró, el pronóstico obtenido mediante suavizamiento exponencial para el periodo 2 es igual al valor real en la serie de tiempo para el periodo 1. Por tanto, como $Y_1 = 17$, para empezar con los cálculos del suavizamiento exponencial se hace $F_2 = 17$. De regreso con los datos de la serie de tiempo, presentados en la tabla 18.1, se encuentra que el valor real para el periodo 2 es $Y_2 = 21$. El error de pronóstico del periodo 2 es $21 - 17 = 4$.

Al continuar con los cálculos del suavizamiento exponencial, con $\alpha = 0.2$ como constante de suavizamiento, se obtiene el pronóstico siguiente para el periodo 3.

$$F_3 = 0.2Y_2 + 0.8F_2 = 0.2(21) + 0.8(17) = 17.8$$

Una vez conocido el valor real para el periodo 3 de la serie de tiempo, $Y_3 = 19$, se puede generar el pronóstico para el periodo 4

$$F_4 = 0.2Y_3 + 0.8F_3 = 0.2(19) + 0.8(17.8) = 18.04$$

Si continúa de esta manera con los cálculos para el suavizamiento exponencial se determinan los pronósticos semanales y los errores semanales de pronóstico como se muestra en la tabla 18.3. Observe que para el periodo 1 no se da ningún pronóstico obtenido mediante suavizamiento exponencial ni tampoco ningún error de pronóstico, ya que no se obtuvo ningún pronóstico. Para la semana 12 se tiene $Y_{12} = 22$ y $F_{12} = 18.48$. ¿Se puede emplear esta información para generar un pronóstico para la semana 13 antes de que se conozca el valor real de la semana 13? Con el modelo de suavizamiento exponencial, se tiene

$$F_{13} = 0.2Y_{12} + 0.8F_{12} = 0.2(22) + 0.8(18.48) = 19.18$$

Por tanto, el pronóstico obtenido mediante suavizamiento exponencial para la semana 13 es 19.18 o 19,180 galones de gasolina. Este pronóstico le será útil a la empresa para la planeación y para la toma de decisiones. La exactitud de este pronóstico no se conocerá sino hasta el final de la semana 13.

En la figura 18.7 se muestra una gráfica con los valores reales y pronosticados de la serie de tiempo. Observe, en especial, cómo los pronósticos suavizan la irregularidad de las fluctuaciones de la serie de tiempos.

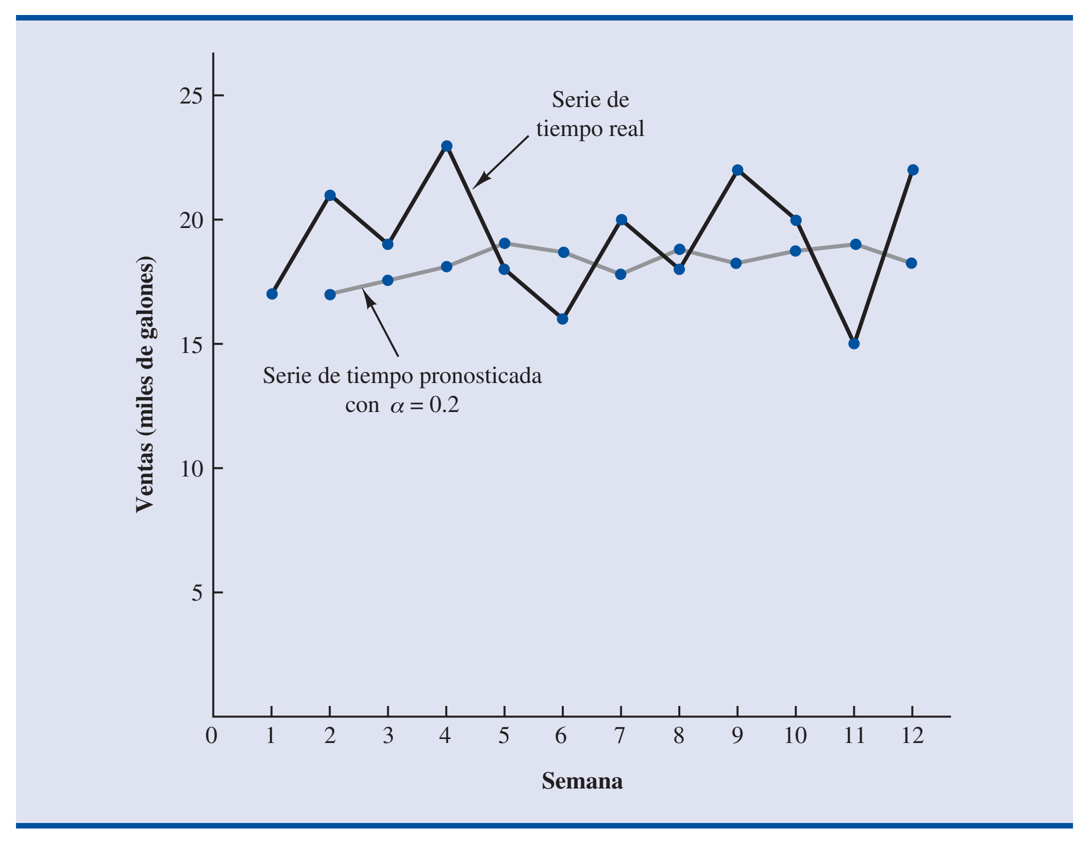
Exactitud del pronóstico En los cálculos anteriores para el suavizamiento exponencial, se empleó como constante de suavizamiento $\alpha = 0.2$. Aunque para α se puede usar cualquier valor en-

TABLA 18.3 RESUMEN DE LOS PRONÓSTICOS OBTENIDOS POR SUAVIZAMIENTO EXPONENCIAL Y DE LOS ERRORES DE PRONÓSTICO PARA LAS VENTAS DE GASOLINA, CON $\alpha = 0.2$ COMO CONSTANTE DE SUAVIZAMIENTO

Semana (t)	Valores en la serie de tiempo (Y_t)	Pronóstico con suavizamiento exponencial (F_t)	Error de pronóstico ($Y_t - F_t$)
1	17		
2	21	17.00	4.00
3	19	17.80	1.20
4	23	18.04	4.96
5	18	19.03	−1.03
6	16	18.83	−2.83
7	20	18.26	1.74
8	18	18.61	−0.61
9	22	18.49	3.51
10	20	19.19	0.81
11	15	19.35	−4.35
12	22	18.48	3.52



FIGURA 18.7 SERIES DE TIEMPO, REAL Y PRONOSTICADA, DE LAS VENTAS DE GASOLINA, CON α = 0.2 COMO CONSTANTE DE SUAVIZAMIENTO



tre 0 y 1, con algunos valores se obtendrá un mejor pronóstico que otros. Una idea de cómo elegir el mejor valor para α se obtiene al revisar el modelo básico de suavizamiento exponencial.

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t$$
$$F_{t+1} = \alpha Y_t + F_t - \alpha F_t$$
$$F_{t+1} = F_t + \alpha(Y_t - F_t) \quad \textbf{(18.3)}$$

Pronóstico para el periodo t — Error de pronóstico en el periodo t

De manera que el nuevo pronóstico F_{t+1} es igual al pronóstico anterior F_t más un ajuste, el cual es igual a α multiplicado por el error de pronóstico más reciente, $Y_t - F_t$. Es decir, el pronóstico para el periodo $t + 1$ se obtiene al ajustar el pronóstico para el periodo t mediante una fracción del error de pronóstico. Si en la serie de tiempo existe una gran variabilidad aleatoria, se prefiere un valor pequeño para la constante de suavizamiento. La razón es que como gran parte del error de pronóstico se debe a la variabilidad aleatoria, no se quiere reaccionar de manera exagerada y ajustar el pronóstico muy rápidamente. En una serie de tiempo con variabilidad aleatoria relativamente pequeña, valores mayores para la constante de suavizamiento permiten ajustar rápidamente los pronósticos cuando ocurren errores de pronóstico, esto permite adaptar los pronósticos, en forma rápida, a las condiciones cambiantes.

El criterio que se usará para determinar el valor adecuado para la constante de suavizamiento α es el mismo que el propuesto para determinar el número de periodos a incluir en el cálculo de los promedios móviles. Es decir, se elige el valor α que minimice el cuadrado medio debido al error (CME). En la tabla 18.4 se presenta, para las ventas de gasolina, un resumen de los cálcu-

TABLA 18.4 CÁLCULO DEL CME DE LOS PRONÓSTICOS PARA LAS VENTAS DE GASOLINA CON $\alpha = 0.2$

Semana (t)	Valores de la serie de tiempos (Y_t)	Pronóstico (F_t)	Error de pronóstico ($Y_t - F_t$)	Cuadrado del error de pronóstico $(Y_t - F_t)^2$
1	17			
2	21	17.00	4.00	16.00
3	19	17.80	1.20	1.44
4	23	18.04	4.96	24.60
5	18	19.03	−1.03	1.06
6	16	18.83	−2.83	8.01
7	20	18.26	1.74	3.03
8	18	18.61	−0.61	0.37
9	22	18.49	3.51	12.32
10	20	19.19	0.81	0.66
11	15	19.35	−4.35	18.92
12	22	18.48	3.52	12.39
			Total	98.80

$$\text{CME} = \frac{98.80}{11} = 8.98$$

los del CME en los pronósticos obtenidos mediante suavizamiento exponencial con $\alpha = 0.2$. Observe que hay un cuadrado del error menos que el número de periodos de tiempo, esto se debe a que no se cuenta con un valor anterior para obtener un pronóstico para el periodo 1. ¿Habrá un valor de α que dé mejores resultados en términos de un valor menor para CME? La manera más sencilla de responder esta pregunta es probar otros valores para α, y después comparar los cuadrados medios de los errores con el valor 8.98 obtenido para el CME con $\alpha = 0.2$.

En la tabla 18.5 se muestran los resultados del suavizamiento exponencial con $\alpha = 0.3$. Como el CME = 9.35, en este conjunto de datos, al emplear como constante de suavizamiento $\alpha = 0.3$ se

TABLA 18.5 CÁLCULO DEL CME PARA LOS PRONÓSTICOS DE LAS VENTAS DE GASOLINA CON $\alpha = 0.3$

Semana (t)	Valores de la serie de tiempo (Y_t)	Pronóstico (F_t)	Error de pronóstico ($Y_t - F_t$)	Cuadrado del error de pronóstico $(Y_t - F_t)^2$
1	17			
2	21	17.00	4.00	16.00
3	19	18.20	0.80	0.64
4	23	18.44	4.56	20.79
5	18	19.81	−1.81	3.28
6	16	19.27	−3.27	10.69
7	20	18.29	1.71	2.92
8	18	18.80	−0.80	0.64
9	22	18.56	3.44	11.83
10	20	19.59	0.41	0.17
11	15	19.71	−4.71	22.18
12	22	18.30	3.70	13.69
			Total	102.83

$$\text{CME} = \frac{102.83}{11} = 9.35$$

obtiene menos exactitud en los pronósticos que si se empleara la constante de suavizamiento $\alpha = 0.2$. Por tanto, se preferirá la constante de suavizamiento original, $\alpha = 0.2$. Al probar con otros valores de α se puede hallar un "buen" valor para la constante de suavizamiento. Este valor puede emplearse en el modelo de suavizamiento exponencial para obtener pronósticos para el futuro. En un momento posterior, después de haber obtenido nuevas observaciones para la serie de tiempo, se vuelven a analizar los datos recolectados de la serie de tiempo y se determina si es necesario modificar la constante de suavizamiento para obtener mejores pronósticos.

NOTAS Y COMENTARIOS

1. Otra medida de la exactitud de los pronósticos es la *desviación absoluta de la media* (DAM). Esta medida es simplemente el promedio de los valores absolutos de todos los errores de pronóstico. Con los errores que se presentan en la tabla 18.2 se obtiene

$$\text{DAM} = \frac{4 + 3 + 4 + 1 + 0 + 4 + 0 + 5 + 3}{9} = 2.67$$

   Una de las principales diferencias entre CME y DAM es que al CME lo influyen mucho más los errores de pronóstico grandes que los errores de pronóstico pequeños (debido a que para el CME los errores se elevan al cuadrado). La elección de la mejor medida para la precisión del pronóstico no es un asunto sencillo. Hasta los expertos en la materia suelen no ponerse de acuerdo respecto a qué medida deba usarse. En este capítulo se usará el CME.
2. Los paquetes de hojas de cálculo son una buena ayuda en la elección de un valor adecuado para α en el suavizamiento exponencial así como en la elección de los pesos en el método de los pesos móviles ponderados. Al tener en la hoja de cálculo los datos de la serie de tiempo y las fórmulas de pronóstico, se pueden probar distintos valores de α (o pesos para el promedio móvil) y elegir el valor o los valores con que se obtenga el menor CME o DAM.

Ejercicios

Métodos



1. Considere la serie de tiempo siguiente.

Semana	1	2	3	4	5	6
Valor	8	13	15	17	16	9

   a. Obtenga un promedio móvil de tres semanas para esta serie de tiempo. ¿Cuál es el pronóstico para la semana 7?
   b. Calcule el CME de este promedio móvil de tres semanas.
   c. Use $\alpha = 0.2$ para calcular los valores de suavizamiento exponencial de esta serie de tiempo. ¿Cuál es el pronóstico para la semana 7?
   d. Compare el pronóstico obtenido con el promedio móvil de tres semanas, con el pronóstico obtenido con el suavizamiento exponencial usando $\alpha = 0.2$. ¿Con cuál se obtiene un mejor pronóstico?
   e. Use 0.4 como constante de suavizamiento, calcule los valores que se obtienen mediante suavizamiento exponencial.

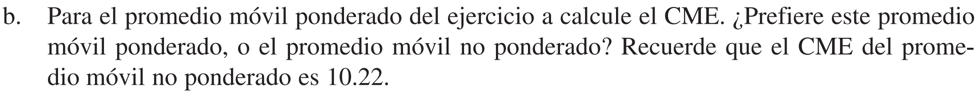




2. Vaya a la serie de tiempo de las ventas de gasolina, presentada en la tabla 18.1.
   a. Para esa serie de tiempo calcule los promedios móviles de cuatro y de cinco semanas.
   b. Calcule el CME de los pronósticos obtenidos con los promedios móviles de cuatro semanas y con los promedios móviles de cinco semanas.
   c. ¿Cuántas semanas es mejor usar para calcular el promedio móvil? Recuerde que el CME del promedio móvil de tres semanas es 10.22.



3. Vaya a la serie de tiempo de las ventas de gasolina, presentada en la tabla 18.1.
   a. Para esa serie de tiempo calcule un promedio móvil ponderado de tres semanas, emplee 1/2 como peso para la observación más reciente, 1/3 para la siguiente observación y 1/6 para la observación más antigua.



b. Para el promedio móvil ponderado del ejercicio a calcule el CME. ¿Prefiere este promedio móvil ponderado, o el promedio móvil no ponderado? Recuerde que el CME del promedio móvil no ponderado es 10.22.

c. Suponga que se permite elegir los pesos, con la única condición de que su suma sea uno. ¿Siempre será posible elegir un conjunto de pesos que hagan que el CME sea menor para un promedio móvil ponderado que para el promedio móvil no ponderado?



Gasoline

4. Para la serie de tiempo de las ventas de gasolina, presentada en la tabla 18.1, dé los pronósticos, emplee $\alpha = 0.1$. Si aplica el criterio del CME, ¿qué constante de suavizamiento será preferible emplear, 0.1 o 0.2?

5. Emplee como constante de suavizamiento $\alpha = 0.2$, la ecuación (18.2) indica que el pronóstico para la semana 13 de las ventas de gasolina, que se presentan en la tabla 18.1, está dado por $F_{13} = 0.2Y_{12} + 0.8F_{12}$, pero como el pronóstico para la semana 12 está dado por $F_{12} = 0.2Y_{11} + 0.8F_{11}$, al combinar estas dos ecuaciones, el pronóstico para la semana 12 se puede escribir como

$$F_{13} = 0.2Y_{12} + 0.8(0.2Y_{11}+0.8F_{11}) = 0.2Y_{12} + 0.16Y_{11} + 0.64F_{11}$$

a. Haga uso de $F_{11} = 0.2Y_{10} + 0.8F_{10}$ (y de manera similar F_{10} y F_9), continúe expandiendo la expresión para F_{13} hasta que quede en función de los valores pasados Y_{12}, Y_{11}, Y_{10}, Y_9, Y_8 y del pronóstico para el periodo 8.

b. Observe los coeficientes o pesos de los valores del pasado Y_{12}, Y_{11}, Y_{10}, Y_9, Y_8; ¿qué puede decir acerca de los pesos del suavizamiento exponencial de los valores pasados al tener un nuevo pronóstico? Compare estos pesos con los pesos en el método del promedio móvil.

Aplicaciones

6. En la empresa Hawkins, los porcentajes de pedidos recibidos a tiempo en los últimos 12 meses son 80, 82, 84, 83, 83, 84, 85, 84, 82, 83, 84 y 83.

a. Compare un pronóstico obtenido con el método de promedios móviles, use promedios de tres meses, con el pronóstico obtenido con el método de suavizamiento exponencial, emplee $\alpha = 0.2$. ¿Con qué método se obtiene un mejor pronóstico?

b. ¿Cuál es el pronóstico para el mes próximo?

7. A continuación se dan las tasas de interés de bonos corporativos triple A en 12 meses consecutivos.

9.5 9.3 9.4 9.6 9.8 9.7 9.8 10.5 9.9 9.7 9.6 9.6

a. Desarrolle promedios móviles de tres y cuatro meses para esta serie de tiempo. ¿Cuál de los dos promedios proporciona el mejor pronóstico? Explique.

b. ¿Cuál es el pronóstico del promedio móvil para el mes próximo?



8. A continuación se presentan los valores (en millones de dólares) de los contratos de construcción en Alabama correspondientes a un periodo de 12 meses.

240 350 230 260 280 320 220 310 240 310 240 230

a. Compare el pronóstico obtenido con promedios móviles de tres meses con el pronóstico obtenido mediante suavizamiento exponencial con $\alpha = 0.2$. ¿Con cuál de los métodos se obtiene un mejor pronóstico?

b. ¿Cuál es el pronóstico para el mes próximo?

9. En la serie de tiempo siguiente se dan las ventas de un determinado producto en los últimos 12 meses.

Mes	Ventas	Mes	Ventas
1	105	7	145
2	135	8	140
3	120	9	100
4	105	10	80
5	90	11	100
6	120	12	110

a. Emplee $\alpha = 0.3$ y calcule los valores que se obtienen para esta serie de tiempo con el método de suavizamiento exponencial.
b. Use la constante de suavizamiento $\alpha = 0.5$ para calcular los valores de suavizamiento exponencial. ¿Con cuál valor, 0.3 o 0.5, de la constante α se obtiene un mejor pronóstico?

10. Los datos siguientes son los valores del índice Commodity Futures Index en 10 semanas: 7.35, 7.40, 7.55, 7.56, 7.60, 7.52, 7.52, 7.70, 7.62 y 7.55.
 a. Dé los valores que se obtienen con el método de suavizamiento exponencial con $\alpha = 0.2$.
 b. Proporcione los valores que se obtienen con el método de suavizamiento exponencial con $\alpha = 0.3$.
 c. ¿Cuál de los dos modelos anteriores proporciona mejores pronósticos? Dé el pronóstico para la semana 11.

11. Los datos siguientes corresponden a la utilización de la capacidad de producción (en porcentajes) en los últimos 15 trimestres.



MfgCap

Trimestre/año	Utilización (%)	Trimestre/año	Utilización (%)
1/2003	82.5	1/2005	78.8
2/2003	81.3	2/2005	78.7
3/2003	81.3	3/2005	78.4
4/2003	79.0	4/2005	80.0
1/2004	76.6	1/2006	80.7
2/2004	78.0	2/2006	80.7
3/2004	78.4	3/2006	80.8
4/2004	78.0		

 a. Para esta serie de tiempo calcule promedios móviles de tres semanas y promedios móviles de cuatro semanas. ¿Con cuál de estos promedios móviles se obtiene un mejor pronóstico para el cuarto trimestre de 2006?
 b. Obtenga pronósticos para el cuarto trimestre de 2006, use $\alpha = 0.4$ y $\alpha = 0.5$. ¿Con qué constantes de suavizamiento se obtiene un mejor pronóstico?
 c. De acuerdo con los resultados de los incisos a y b, ¿con qué método —promedios móviles o suavizamiento exponencial— se obtiene un mejor pronóstico? Explique.



Proyección de tendencia

En esta sección se muestra cómo hacer pronósticos para una serie de tiempo que a largo plazo tenga una tendencia lineal. El tipo de series de tiempo al que se aplica el método de la proyección de tendencia son las series de tiempo que muestran un aumento o disminución consistente a lo largo del tiempo; como estas series de tiempo no son estables, no se pueden utilizar los métodos de suavizamiento descritos en la sección anterior.

Considere la serie de tiempo formada por las ventas de bicicletas de un determinado fabricante en los últimos 10 años, la cual se muestra en la tabla 18.6 y en la figura 18.8. Observe que en el año 1 se vendieron 21 600 bicicletas, en el año 122 900 en el año 2, etc. En el año 10, el año más reciente, se vendieron 31 400 bicicletas. Aunque en la figura 18.8 se observa que en los últimos 10 años ha habido algunos ascensos y algunos descensos, esta serie de tiempo parece mostrar una tendencia general de aumento o creciente.

TABLA 18.6 SERIE DE TIEMPO DE LA VENTA DE BICICLETAS

Año (t)	Ventas (en miles) (Y_t)
1	21.6
2	22.9
3	25.5
4	21.9
5	23.9
6	27.5
7	31.5
8	29.7
9	28.6
10	31.4

No puede esperar que el componente de tendencia de una serie de tiempo siga todos y cada uno de los movimientos de ascenso y descenso. Lo que se espera es que el componente de tendencia se refleje en el desplazamiento gradual —en este caso, un crecimiento— de los valores de la serie de tiempo. Después de observar los datos de la serie de tiempo que se presentan en la tabla 18.6 y su gráfica que aparece en la figura 18.8, se puede estar de acuerdo en que una descripción razonable de los movimientos a largo plazo de esta serie es una tendencia lineal como la que se muestra en la figura 18.9.

FIGURA 18.8 SERIE DE TIEMPO DE LAS VENTAS DE BICICLETAS

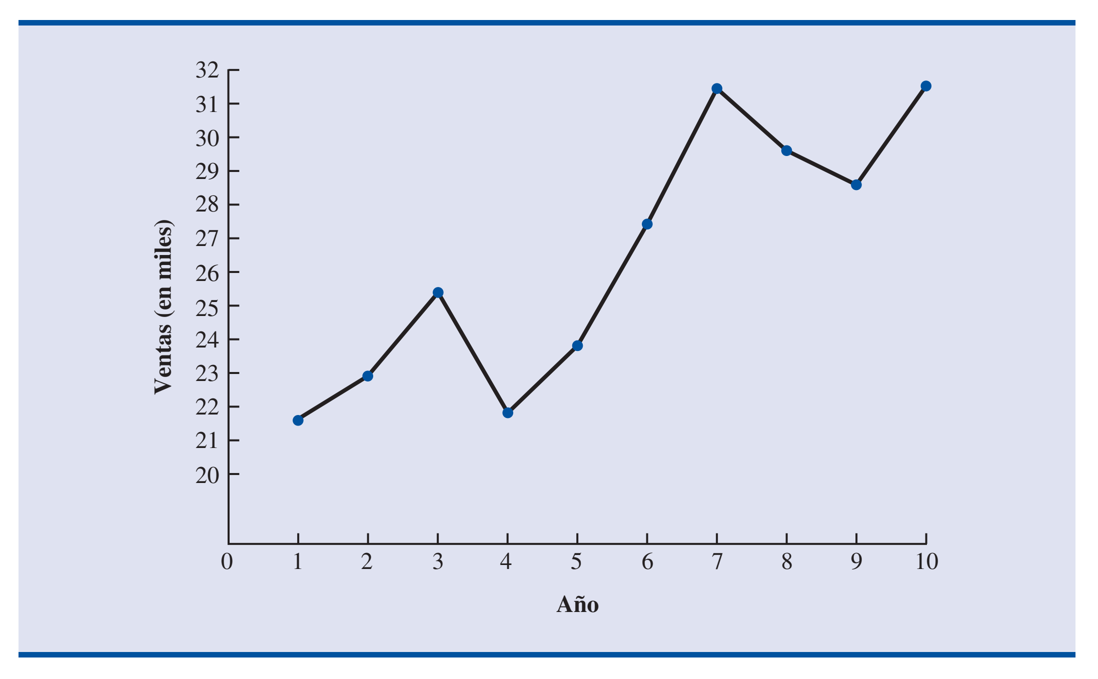
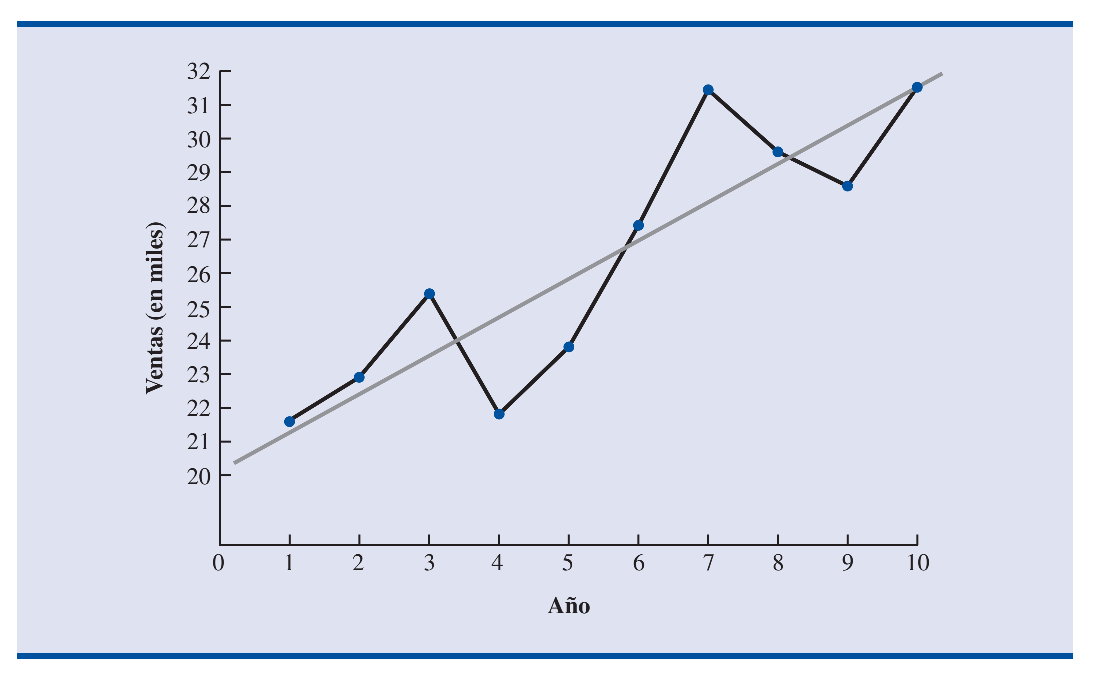


FIGURA 18.9 TENDENCIA REPRESENTADA POR UNA FUNCIÓN LINEAL PARA EL CASO DE LAS VENTAS DE BICICLETAS





Para ilustrar los cálculos a realizar al emplear el análisis de regresión para identificar una tendencia lineal, se aprovecharán los datos de las ventas de bicicletas. Recuerde que en el capítulo 14, cuando se estudió la regresión lineal simple, se vio cómo usar el método de mínimos cuadrados para hallar la mejor relación lineal entre dos variables. Para obtener la línea de tendencia para la serie de tiempo de las ventas de bicicletas se empleará esa misma metodología. Es decir, se usará el análisis de regresión para estimar la relación entre tiempo y volumen de ventas.

En el capítulo 14 la ecuación de regresión estimada que describía una relación lineal entre una variable independiente x y una variable dependiente y se expresó como

$$\hat{y} = b_0 + b_1 x \quad \textbf{(18.4)}$$

Para hacer énfasis en el hecho de que al hacer un pronóstico la variable independiente es el tiempo, en la ecuación (18.4) se usará t en lugar de x; además, en lugar de $\hat{y}$ se usará T_t. Por tanto, en el caso de una tendencia lineal, el volumen de ventas estimado que se expresa en función del tiempo se escribe

ECUACIÓN DE TENDENCIA LINEAL

$$T_t = b_0 + b_1 t \quad \textbf{(18.5)}$$

donde

T_t = valor de tendencia de la serie de tiempo en el periodo t
b_0 = intersección de la línea de tendencia
b_1 = pendiente de la línea de tendencia
t = tiempo

En la ecuación (18.5), se hará $t = 1$ para el tiempo de la primera observación de la serie de tiempo, $t = 2$ para el tiempo de la segunda observación y así sucesivamente. Observe que en la serie de tiempo de las ventas de bicicletas, $t = 1$ corresponde al valor más antiguo de la serie, y $t = 10$ corresponde al dato del año más reciente. A continuación se presentan las fórmulas para calcular los coeficientes de regresión estimados (b_1 y b_0) de la ecuación (18.5).

CÁLCULO DE LA PENDIENTE (b_1) Y DE LA INTERSECCIÓN (b_0)

$$b_1 = \frac{\Sigma t Y_t - (\Sigma t \Sigma Y_t)/n}{\Sigma t^2 - (\Sigma t)^2/n} \quad \textbf{(18.6)}$$

$$b_0 = \bar{Y} - b_1 \bar{t} \quad \textbf{(18.7)}$$

donde

Y_t = valor en la serie de tiempo en el periodo t
n = número de periodos
$\bar{Y}$ = valor promedio de la serie de tiempo; es decir, $\bar{Y} = \Sigma Y_t / n$
$\bar{t}$ = valor promedio de t, es decir, $\bar{t} = \Sigma t / n$

Al emplear las ecuaciones (18.6) y (18.7) y los datos de las ventas de bicicletas presentados en la tabla 18.6, b_0 y b_1 se calculan como sigue.

t	Y_t	tY_t	t^2
1	21.6	21.6	1
2	22.9	45.8	4
3	25.5	76.5	9
4	21.9	87.6	16
5	23.9	119.5	25
6	27.5	165.0	36

	t	Y_t	tY_t	t^2
	7	31.5	220.5	49
	8	29.7	237.6	64
	9	28.6	257.4	81
	10	31.4	314.0	100
Totales	55	264.5	1545.5	385

$$\bar{t} = \frac{55}{10} = 5.5$$

$$\bar{Y} = \frac{264.5}{10} = 26.45$$

$$b_1 = \frac{1545.5 - (55)(264.5)/10}{385 - (55)^2/10} = 1.10$$

$$b_0 = 26.45 - 1.10(5.5) = 20.4$$

Por tanto,

$$T_t = 20.4 + 1.1t \qquad \textbf{(18.8)}$$

es la expresión del componente de tendencia lineal en la serie de tiempo de las ventas de bicicletas.

Antes de usar la ecuación de tendencia para obtener un pronóstico, debe realizar una prueba de significancia estadística (ver capítulo 14). En la práctica, esta prueba debe ser parte rutinaria del ajuste de la línea de tendencia.

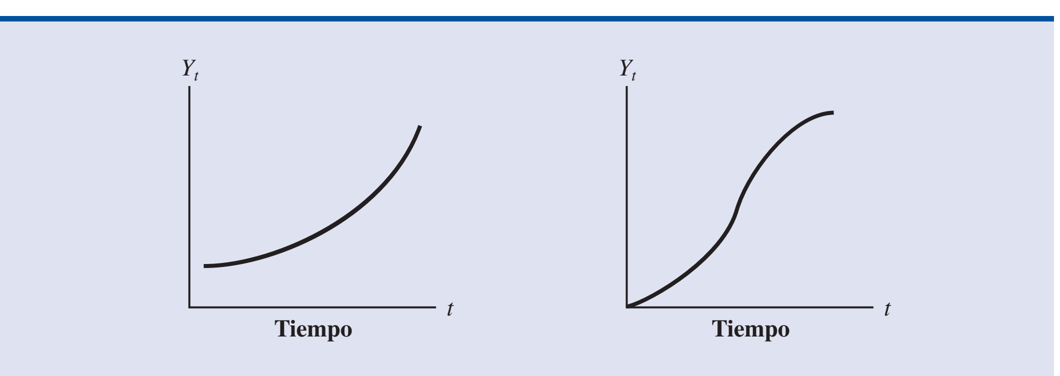
Como la pendiente es 1.1, esto indica que en los pasados 10 años la empresa tuvo un crecimiento promedio en ventas de 1 100 unidades por año. Si se supone que la tendencia en ventas de los últimos 10 años es un buen indicador del futuro, entonces se emplea la ecuación 18.8 para proyectar el componente de tendencia de la serie de tiempo. Por ejemplo, al sustituir en la ecuación (18.8) $t = 11$, se obtiene la proyección de tendencia para el año próximo, T_{11}.

$$T_{11} = 20.4 + 1.1(11) = 32.5$$

Por tanto, si emplea únicamente el componente de tendencia se pronostica que, el año próximo, las ventas serán de 32 500 bicicletas.

Para modelar una tendencia suele usarse el modelo de función lineal, sin embargo, como ya se vio, algunas veces las series de tiempo tienen tendencias curvilíneas, o no lineales, similares a las que se muestran en la figura 18.10. En el capítulo 16 se vio cómo usar el análisis de regresión para modelar relaciones curvilíneas como la presentada en la gráfica A de la figura 18.10. En libros más avanzados se ve detalladamente cómo obtener modelos de regresión como el mostrado en la gráfica B de la figura 18.10.

FIGURA 18.10 FORMAS POSIBLES PARA UN PATRÓN DE TENDENCIA NO LINEAL



Gráfica A. Tendencia exponencial **Gráfica B. Curva de crecimiento de Gompertz**

Ejercicios

Métodos



12. Considere la serie de tiempo siguiente.

t	1	2	3	4	5
Y_t	6	11	9	14	15

Obtenga una ecuación para el componente de tendencia lineal de esta serie. Dé el pronóstico para $t = 6$.

13. Considere la serie de tiempo siguiente.

t	1	2	3	4	5	6
Y_t	205	202	195	190	191	188

Obtenga una ecuación para el componente de tendencia lineal de esta serie. Dé el pronóstico para $t = 7$.

Aplicaciones

Autoexamen

14. Se presentan los datos de matrícula (en miles) en una universidad en los últimos seis años.

Año	1	2	3	4	5	6
Matrícula	20.5	20.2	19.5	19.0	19.1	18.8

Obtenga la ecuación para el componente de tendencia lineal de esta serie. Haga un comentario sobre lo que pasa con la matrícula en esta institución.

15. En la tabla siguiente se dan las cifras correspondientes a los últimos siete años, de asistencia promedio a los juegos de futbol, en casa, de una universidad. Obtenga la ecuación para el componente de tendencia lineal de esta serie de tiempo.

Año	Asistencia
1	28 000
2	30 000
3	31 500
4	30 400
5	30 500
6	32 200
7	30 800

16. De las ventas de automóviles en los últimos 10 años en B.J. Scott Motors, Inc., se obtuvo la serie de tiempo siguiente.

Año	Ventas	Año	Ventas
1	400	6	260
2	390	7	300
3	320	8	320
4	340	9	340
5	270	10	370

Grafique esta serie de tiempo y haga un comentario sobre si es adecuado suponer una tendencia lineal. ¿Qué tipo de forma funcional considera más adecuada para el patrón de tendencia de esta serie?

17. Al presidente de una pequeña fábrica le preocupa el aumento continuo que ha habido en los costos de fabricación en los últimos años. Las cifras siguientes constituyen una serie de tiempo del costo por unidad de los principales productos de esta empresa en los últimos ocho años.

Año	Costo/unidad ($)	Año	Costo/unidad ($)
1	20.00	5	26.60
2	24.50	6	30.00
3	28.20	7	31.00
4	27.50	8	36.00

a. Presente una gráfica de esta serie de tiempo. ¿Parece existir una tendencia lineal?
b. Obtenga la ecuación del componente de tendencia lineal de esta serie de tiempo. ¿Cuál es el incremento anual promedio en el costo que ha habido en la empresa?

18. Los datos siguientes presentan el porcentaje de estadounidenses rurales, urbanos y suburbanos que tienen en casa una conexión de alta velocidad (Pew Internet Rural Broadband Internet Use Memo, febrero de 2006).



Broadband

Año	Rural	Urbano	Suburbano
2001	3	9	9
2002	6	18	17
2003	9	21	23
2004	16	29	29
2005	24	38	40

a. Para cada grupo, obtenga una ecuación de tendencia lineal.
b. Utilice la ecuación de tendencia obtenida en el inciso a para comparar las tasas de crecimiento de los tres grupos.
c. Use la ecuación de tendencia de cada grupo para obtener un pronóstico para el año 2006.

19. En los datos siguientes se observan los promedios en las cuentas por teléfonos celulares (*The New York Times Almanac,* 2006)

Año	Cantidad ($)
1998	39.43
1999	41.24
2000	45.27
2001	47.37
2002	48.40
2003	49.91

a. Grafique esta serie de tiempo. ¿Observa alguna tendencia?
b. Para esta serie de tiempo obtenga una ecuación de tendencia lineal.
c. Use la ecuación de tendencia para estimar la cuenta promedio mensual en 2004.

20. A continuación se presentan los datos del ingreso bruto (en millones de dólares) de las aerolíneas regionales en un periodo de 10 años.

Año	Ingreso	Año	Ingreso
1	2428	6	4264
2	2951	7	4738
3	3533	8	4460
4	3618	9	5318
5	3616	10	6915

a. Para esta serie de tiempo, obtenga una ecuación de tendencia lineal. Haga un comentario sobre lo que revela esta ecuación acerca del ingreso bruto de las aerolíneas en los últimos 10 años.
b. Pronostique los ingresos brutos en los años 11 y 12.

21. FRED® (Federal Reserve Economic Data), una base de datos que con más de 3 000 series de tiempo económicas contiene datos históricos sobre los tipos de cambio. Los datos siguientes corresponden al tipo de cambio entre Estados Unidos y Canadá (http://research.stlouisfed.org/fred2/). Las unidades en esta tasa de cambio son cantidad de dólares canadienses por un dólar estadounidense.

Fecha	Cambio
Abril 2005	1.2359
Mayo 2005	1.2555
Junio 2005	1.2402
Julio 2005	1.2229
Agosto 2005	1.2043
Septiembre 2005	1.1777
Octubre 2005	1.1774
Noviembre 2005	1.1815
Diciembre 2005	1.1615
Enero 2006	1.1572

a. Grafique esta serie de tiempo. ¿Está presente una tendencia lineal?
b. Obtenga la ecuación para el componente de tendencia lineal de esta serie de tiempo.
c. Use la ecuación de tendencia para pronosticar el tipo de cambio en febrero 2006.
d. ¿Usaría con confianza esta ecuación para pronosticar el tipo de cambio en julio de 2006?



Componentes de tendencia y estacionales

Ya se mostró cómo hacer pronósticos para una serie de tiempo que tiene un componente de tendencia. En esta sección se amplía este estudio y se muestra cómo hacer pronósticos para una serie de tiempo que tiene tanto un componente de tendencia como un componente estacional.

En muchas situaciones relacionadas con las actividades comerciales o económicas es necesario hacer comparaciones entre los periodos. Informaciones en el sentido de que ha habido un aumento de 2% en el desempleo, en comparación con el último mes; o un crecimiento de 5% en la producción de acero, en comparación con el mes anterior, o una disminución de 3% en el consumo en la energía eléctrica, en comparación con el mes anterior, suelen ser de interés. Sin embargo, hay que tener cuidado con este tipo de información, ya que si existe una influencia estacional, esta información puede entenderse de manera equivocada. Por ejemplo, el hecho de que el consumo de energía haya disminuido 3% de agosto a septiembre puede deberse únicamente al efecto estacional, al menor uso que se hace del aire acondicionado, y no a una disminución a largo plazo en el consumo de energía eléctrica. En realidad, una vez hecho el ajuste estacional, es posible que incluso se encuentre que el consumo de energía eléctrica haya aumentado.

A la eliminación del efecto estacional de una serie de tiempo se le conoce como desestacionalización de la serie de tiempo. La desestacionalización permite que las comparaciones de un periodo con otro sean útiles y ayuda a identificar la existencia de tendencias. El método que se expondrá aquí es adecuado cuando existen únicamente efectos estacionales o cuando existen tanto efectos estacionales como de tendencia. El primer paso es calcular los índices estacionales y usarlos para desestacionalizar los datos. Después, si en los datos desestacionalizados se observa alguna tendencia, se estima el componente de tendencia aplicando el análisis de regresión a la serie desestacionalizada.

Modelo multiplicativo

Se supone que en la serie de tiempo además del componente de tendencia (T) y del componente estacional (S) existe un componente irregular (I). El componente irregular corresponde a todos los efectos aleatorios que puede haber sobre la serie de tiempo, es decir, los efectos que no pueden ser

explicados por los componentes de tendencia y estacional. Para identificar los componentes de tendencia, estacional e irregular correspondientes a un tiempo *t* se usarán T_t, S_t e I_t, respectivamente, y se supondrá que el valor correspondiente de la serie de tiempo, que se denotará Y_t, se puede determinar mediante el siguiente **modelo multiplicativo para una serie de tiempo**.

$$Y_t = T_t \times S_t \times I_t \tag{18.9}$$

En este modelo, T_t es la tendencia, y se mide en las unidades del artículo que se pronostica. En cambio, los componentes S_t e I_t se miden en términos relativos, donde valores mayores a 1.00 indican efectos superiores a la tendencia y valores menores a 1.00 indican valores inferiores a la tendencia.

Para ilustrar el uso del modelo multiplicativo que comprende los componentes de tendencia, estacional e irregular se emplearán los datos trimestrales que se muestran en la tabla 18.7 y en la figura 18.11. Estos son los datos que corresponden a las ventas de unidades de televisores (en miles) que ha vendido un determinado fabricante en los últimos cuatro años. Primero se mostrará cómo identificar el componente estacional en la serie de tiempo.

TABLA 18.7

DATOS TRIMESTRALES DE LA VENTA DE TELEVISORES

Año	Tri-mestre	Ventas (en miles)
1	1	4.8
	2	4.1
	3	6.0
	4	6.5
2	1	5.8
	2	5.2
	3	6.8
	4	7.4
3	1	6.0
	2	5.6
	3	7.5
	4	7.8
4	1	6.3
	2	5.9
	3	8.0
	4	8.4

Cálculo de los índices estacionales

En la figura 18.11 se observa que cada año las ventas disminuyen en el segundo trimestre, y aumentan en el tercero y cuarto trimestres. Por tanto, se concluye que en la venta de estos televisores hay un comportamiento estacional. Para identificar el componente estacional de cada trimestre, se empieza por calcular un promedio móvil para separar los componentes estacional e irregular, S_t e I_t, del componente de tendencia T_t.

Para calcular cada promedio se usan datos de todo un año. Como se trabaja con una serie trimestral, para calcular cada promedio móvil se usan cuatro datos. El cálculo del promedio móvil de los cuatro primeros trimestres de ventas de televisores es

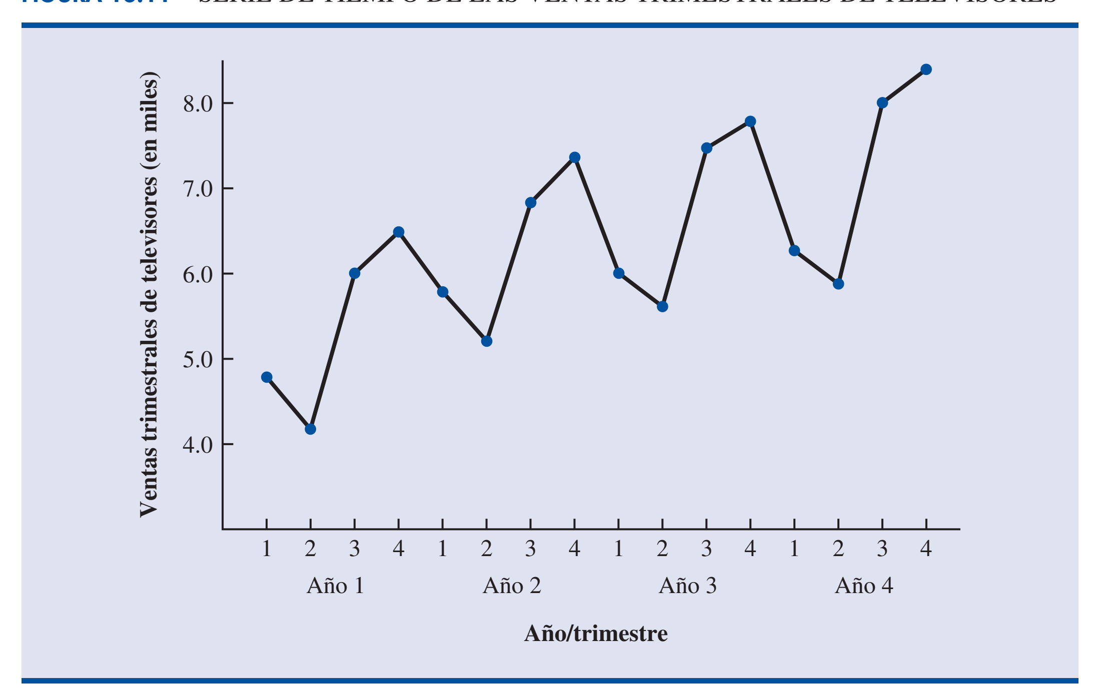


TVSales

$$\text{Primer promedio móvil} = \frac{4.8 + 4.1 + 6.0 + 6.5}{4} = \frac{21.4}{4} = 5.35$$

Observe que el promedio móvil de los primeros cuatro trimestres da el promedio trimestral de las ventas en el año 1 de la serie de tiempo. Para continuar con el cálculo de los promedios mó-

FIGURA 18.11 SERIE DE TIEMPO DE LAS VENTAS TRIMESTRALES DE TELEVISORES



viles, se agrega el valor 5.8 correspondiente al primer trimestre del año 2 y se elimina el 4.8 correspondiente al primer trimestre del año 1. De esta manera el segundo promedio móvil es

$$\text{Segundo promedio móvil} = \frac{4.1 + 6.0 + 6.5 + 5.8}{4} = \frac{22.4}{4} = 5.60$$

De manera similar, el tercer promedio móvil es (6.0 + 6.5 + 5.8 + 5.2)/4 = 5.875.

Antes de continuar el cálculo de los promedios móviles de toda la serie, regrese al primer promedio móvil, que es 5.35. El valor 5.35 representa un promedio trimestral del volumen de ventas (a lo largo de todas las estaciones) en el año 1. Parece razonable asociar el valor 5.35 con el trimestre "central" de los promedios empleados en el cálculo del promedio móvil. Sin embargo, como en cada promedio móvil intervienen cuatro trimestres, no hay un trimestre central. El 5.35 corresponderá a la segunda mitad del segundo trimestre y a la primera mitad del tercer trimestre. De manera similar, al pasar al valor del siguiente promedio móvil, que es 5.60, el trimestre central corresponderá a la segunda mitad del tercer trimestre y a la primera mitad del cuarto trimestre.

Recuerde que la razón por la que se calculan los promedios móviles es para aislar los componentes combinados estacional e irregular. Sin embargo, los valores de los promedios móviles no corresponden precisamente a los trimestres originales de la serie de tiempo. Esta dificultad se puede resolver con el uso de los puntos medios entre los valores sucesivos de los promedios móviles. Por ejemplo, si 5.35 corresponde a la primera mitas del tercer trimestre y 5.60 corresponde a la última mitad del tercer trimestre, se puede usar (5.35 + 5.60)/2 = 5.475 como valor correspondiente al promedio móvil del tercer trimestre. De manera similar, se asocia el valor del promedio móvil (5.60 + 5.875)/2 = 5.738 con el cuarto trimestre. Lo que se obtiene es un *promedio móvil centrado*. En la tabla 18.8 se muestra el resumen completo de los promedios móviles para los datos de las ventas de televisores.

Si en un promedio móvil el número de datos es non, el punto medio corresponderá a uno de los periodos de la serie de tiempo. En tales casos, no es necesario centrar los promedios móviles para hacerlos corresponder a un determinado periodo como se hizo con los datos de la tabla 18.8.

¿Qué información se obtiene, de los promedios móviles centrados de la tabla 18.8, acerca de esta serie de tiempo? En la figura 18.12 se grafican los valores reales de la serie de tiempo y los valores de los promedios móviles centrados. Observe, en particular, cómo los promedios móviles centrados tienden a "suavizar" tanto las fluctuaciones estacionales como las fluctuaciones irregulares de la serie de tiempo. En los promedios móviles calculados con cuatro trimestres no aparecen las fluctuaciones debidas a la influencia estacional ya que este efecto es eliminado con los promedios. Cada promedio móvil centrado representa el valor de la serie de tiempo si no existieran las influencias estacional irregular.

Al dividir cada observación de la serie de tiempo entre su correspondiente promedio móvil centrado se identifica el efecto estacional irregular sobre la serie de tiempo. Por ejemplo, para el tercer trimestre del año 1, 6.0/5.475 = 1.096 es el valor combinado estacional irregular. En la tabla 18.9 se dan los valores estacionales irregulares de toda la serie de tiempo.

Respecto del tercer trimestre se observa que los valores de éste en los años 1, 2, y 3 son 1.096, 1.075 y 1.109, respectivamente. De manera que en todos los casos, el valor estacional irregular parece tener una influencia superior a la promedio sobre el tercer trimestre. Dado que las fluctuaciones que se observan año con año en el valor estacional irregular son atribuibles principalmente al componente irregular, pueden promediarse esos valores para eliminar la influencia irregular y obtener una estimación de la influencia estacional del tercer trimestre.

$$\text{Efecto estacional del tercer trimestre} = \frac{1.096 + 1.075 + 1.109}{3} = 1.09$$

Al número 1.09 se le conoce como *índice estacional* del tercer trimestre. En la tabla 18.10 se resumen los cálculos para obtener los índices estacionales de la serie de tiempo de las ventas de te-

TABLA 18.8 PROMEDIOS MÓVILES CENTRADOS DE LA SERIE DE TIEMPO DE LAS VENTAS DE TELEVISORES

Año	Trimestre	Ventas (en miles)	Promedio móvil de cuatro trimestres	Promedios móviles centrados
1	1	4.8		
	2	4.1		
			5.350	
	3	6.0		5.475
			5.600	
	4	6.5		5.738
			5.875	
2	1	5.8		5.975
			6.075	
	2	5.2		6.188
			6.300	
	3	6.8		6.325
			6.350	
	4	7.4		6.400
			6.450	
3	1	6.0		6.538
			6.625	
	2	5.6		6.675
			6.725	
	3	7.5		6.763
			6.800	
	4	7.8		6.838
			6.875	
4	1	6.3		6.938
			7.000	
	2	5.9		7.075
			7.150	
	3	8.0		
	4	8.4		

levisores. Por tanto, los índices estacionales de los cuatro trimestres son: primer trimestre, 0.93; segundo trimestre, 0.84; tercer trimestre, 1.09, y cuarto trimestre 1.14.

Si observa los valores de la tabla 18.10 obtiene una interpretación del componente estacional en las ventas de televisores. En el cuarto trimestre es cuando se tienen las mejores ventas, que son 14% superiores al promedio de las ventas trimestrales. El peor trimestre de ventas es el segundo trimestre; su índice estacional es 0.84, lo que indica que en este trimestre las ventas son 16% inferiores a las ventas promedio de los cuatro trimestres. El componente estacional corresponde con claridad a lo que la intuición prevé: en el cuarto trimestre, debido a la llegada del invierno y a la disminución de las actividades al aire libre (en Estados Unidos) aumenta el interés por ver televisión y con ello las compras de televisores. Las bajas ventas en el segundo trimestre reflejan la disminución del interés por ver televisión debido a la primavera y a las actividades de preparación para el verano, de los compradores potenciales.

Para obtener el índice estacional suele ser necesario un último ajuste. En el modelo multiplicativo se requiere que el índice estacional promedio sea igual a 1.00, de manera que la suma de los cuatro índices estacionales, que se presentan en la figura 18.10, debe ser igual a 4.00. En otras palabras, el efecto estacional debe compensarse a lo largo del año. En el ejemplo visto aquí, el

FIGURA 18.12 SERIE DE TIEMPO DE LAS VENTAS TRIMESTRALES DE TELEVISORES Y PROMEDIOS MÓVILES CENTRADOS

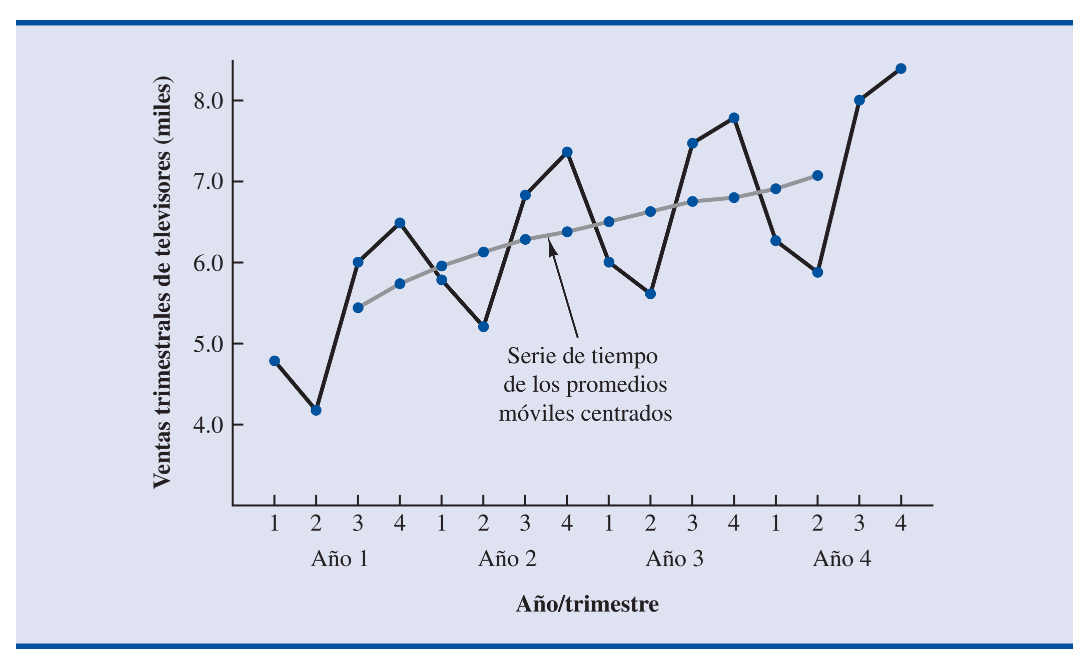


TABLA 18.9 VALORES ESTACIONALES IRREGULARES DE LA SERIE DE TIEMPO DE LAS VENTAS DE TELEVISORES

Año	Trimestre	Ventas (en miles)	Promedio móvil centrada	Valor estacional irregular
1	1	4.8		
	2	4.1		
	3	6.0	5.475	1.096
	4	6.5	5.738	1.133
2	1	5.8	5.975	0.971
	2	5.2	6.188	0.840
	3	6.8	6.325	1.075
	4	7.4	6.400	1.156
3	1	6.0	6.538	0.918
	2	5.6	6.675	0.839
	3	7.5	6.763	1.109
	4	7.8	6.838	1.141
4	1	6.3	6.938	0.908
	2	5.9	7.075	0.834
	3	8.0		
	4	8.4		

TABLA 18.10 CÁLCULO DE LOS ÍNDICES ESTACIONALES DE LA SERIE DE TIEMPO DE LAS VENTAS DE TELEVISORES

Trimestre	Valor del componente estacional irregular (S_tI_t)	Índice estacional (S_t)
1	0.971, 0.918, 0.908	0.93
2	0.840, 0.839, 0.834	0.84
3	1.096, 1.075, 1.109	1.09
4	1.133, 1.156, 1.141	1.14

promedio de los índices estacionales es igual a 1.00 y por tanto, no es necesario hacer ningún ajuste. En otros casos puede ser necesario hacer un ligero ajuste. Este ajuste se hace al multiplicar cada índice estacional por el número de estaciones dividido entre la suma de los índices estacionales no ajustados. Por ejemplo, cuando se tienen datos trimestrales se multiplica cada índice estacional por 4/(suma de los índices estacionales no ajustados). En algunos de los ejercicios se necesitará hacer este ajuste para obtener el índice estacional adecuado.

Desestacionalización de una serie de tiempo

Cuando se tienen datos desestacionalizados, tiene sentido comparar las ventas de periodos consecutivos. Si se tienen datos que no han sido desestacionalizados, comparaciones útiles pueden obtenerse al contrastar las ventas del periodo presente con las ventas del mismo periodo en el año anterior.

El objeto de hallar los índices estacionales es eliminar de la serie de tiempo los efectos estacionales. A este proceso se le conoce como *desestacionalización* de la serie de tiempo. En publicaciones como *Survey of Current Business, The Wall Street Journal* y *BusinessWeek* suelen publicarse series de tiempo económicas ajustadas a las variaciones estacionales (**series de tiempo desestacionalizadas**). Si emplea la noción de modelo multiplicativo, tiene

$$Y_t = T_t \times S_t \times I_t$$

Para eliminar de una serie de tiempo el efecto estacional, se divide cada observación de la serie de tiempo entre su índice estacional correspondiente. En la tabla 18.11 se presenta la serie de tiempo desestacionalizada de las ventas de televisores. En la figura 18.13 se presenta una gráfica de la serie de tiempo desestacionalizada de las ventas de televisores.

Uso de una serie de tiempo desestacionalizada para la identificación de tendencias

A pesar de que en la figura 18.13 se observan algunos movimientos aleatorios, hacia arriba o hacia abajo, a lo largo de los últimos 16 trimestres, la serie de tiempo parece tener una tendencia lineal ascendente. Para identificar esta tendencia, se usa el mismo procedimiento que en la sección anterior; en este caso los datos son ventas trimestrales desestacionalizadas. Por tanto, en una tendencia lineal, el volumen estimado de las ventas, expresado en función del tiempo es

$$T_t = b_0 + b_1 t$$

donde

T_t = valor de la tendencia en la venta de televisores en el periodo t
b_0 = intersección de la línea de tendencia con el eje y
b_1 = pendiente de la línea de tendencia

Como antes, $t = 1$ corresponde al tiempo de la primera observación en la serie de tiempo, $t = 2$ corresponde al tiempo de la segunda observación y así sucesivamente. Así, en la serie de tiempo

TABLA 18.11 VALORES DESESTACIONALIZADOS DE LA SERIE DE TIEMPO DE LAS VENTAS DE TELEVISORES

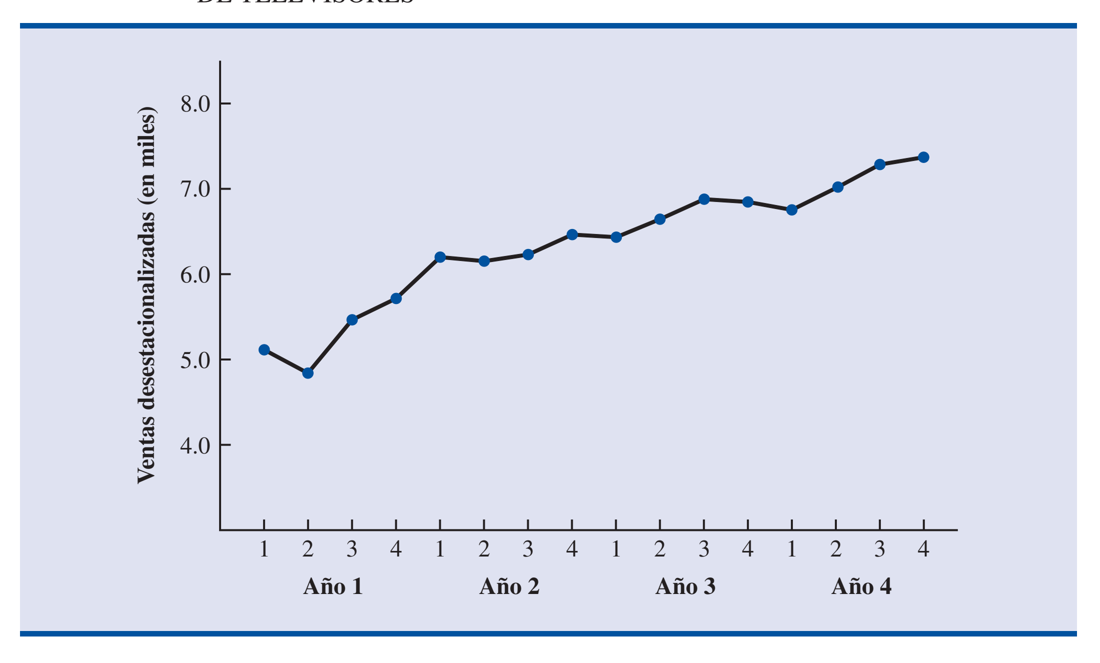
Año	Trimestre	Ventas (en miles) (Y_t)	Índice estacional (S_t)	Ventas desestacionalizadas ($Y_t/S_t = T_tI_t$)
1	1	4.8	.93	5.16
	2	4.1	.84	4.88
	3	6.0	1.09	5.50
	4	6.5	1.14	5.70
2	1	5.8	.93	6.24
	2	5.2	.84	6.19
	3	6.8	1.09	6.24
	4	7.4	1.14	6.49
3	1	6.0	.93	6.45
	2	5.6	.84	6.67
	3	7.5	1.09	6.88
	4	7.8	1.14	6.84
4	1	6.3	.93	6.77
	2	5.9	.84	7.02
	3	8.0	1.09	7.34
	4	8.4	1.14	7.37

de las ventas de televisores, $t = 1$ corresponde al valor de las ventas desestacionalizadas del primer trimestre y $t = 16$ corresponde al valor de las ventas desestacionalizadas del semestre más reciente. A continuación se presentan las fórmulas para calcular los valores de b_0 y b_1.

$$b_1 = \frac{\Sigma tY_t - (\Sigma t \Sigma Y_t)/n}{\Sigma t^2 - (\Sigma t)^2/n}$$

$$b_0 = \bar{Y} - b_1\bar{t}$$

FIGURA 18.13 SERIE DE TIEMPO DESESTACIONALIZADA DE LAS VENTAS DE TELEVISORES



Sin embargo, observe, que ahora Y_t se refiere al valor de la serie de tiempo desestacionalizada en el tiempo t y no al valor de la serie de tiempo original. Al usar las ecuaciones para obtener b_0 y b_1 y los datos de las ventas desestacionalizadas presentados en la tabla 18.11, se tienen los cálculos siguientes.

	t	Y_t (desestacionalizada)	tY_t	t^2
	1	5.16	5.16	1
	2	4.88	9.76	4
	3	5.50	16.50	9
	4	5.70	22.80	16
	5	6.24	31.20	25
	6	6.19	37.14	36
	7	6.24	43.68	49
	8	6.49	51.92	64
	9	6.45	58.05	81
	10	6.67	66.70	100
	11	6.88	75.68	121
	12	6.84	82.08	144
	13	6.77	88.01	169
	14	7.02	98.28	196
	15	7.34	110.10	225
	16	7.37	117.92	256
Totales	136	101.74	914.98	1496

donde

$$\bar{t} = \frac{136}{16} = 8.5$$

$$\bar{Y} = \frac{101.74}{16} = 6.359$$

$$b_1 = \frac{914.98 - (136)(101.74)/16}{1496 - (136)^2/16} = 0.148$$

$$b_0 = 6.359 - 0.148(8.5) = 5.101$$

Por tanto,

$$T_t = 5.101 + 0.148t$$

es la expresión para el componente de tendencia lineal de la serie de tiempo desestacionalizada.

La pendiente, 0.148, indica que en los 16 trimestres pasados, el crecimiento promedio de las ventas desestacionalizadas de la empresa fue de 148 televisores por trimestre. Si se supone que la tendencia en los datos de ventas de los últimos 16 trimestres es un indicador razonablemente bueno del futuro, esta ecuación puede usarse para proyectar el componente de tendencia de la serie de tiempo desestacionalizada a los trimestres futuros. Por ejemplo, si sustituye en esta ecuación $t = 17$, se obtiene la proyección de la tendencia desestacionalizada para el trimestre siguiente, T_{17}.

$$T_{17} = 5.101 + 0.148(17) = 7.617$$

TABLA 18.12 PRONÓSTICOS TRIMESTRALES PARA LA SERIE DE TIEMPO DE LAS VENTAS DE TELEVISORES

Año	Trimestre	Pronóstico para la tendencia desestacionalizada	Índice estacional (ver tabla 18.11)	Pronóstico trimestral
5	1	7617	0.93	(7617)(0.93) = 7084
	2	7765	0.84	(7765)(0.84) = 6523
	3	7913	1.09	(7913)(1.09) = 8625
	4	8061	1.14	(8061)(1.14) = 9190

Por tanto, el componente de tendencia da un pronóstico de ventas desestacionalizadas de 7 617 televisores para el siguiente trimestre. De manera similar, el componente de tendencia produce pronósticos de ventas desestacionalizadas de 7 765, 7 913 y 8 061 televisores para los trimestres 18, 19 y 20, respectivamente.

Ajustes estacionales

El último paso para obtener un pronóstico cuando existen tanto un componente estacional como un componente de tendencia, es usar el índice estacional para ajustar la proyección de tendencia desestacionalizada. De regreso con el ejemplo de las ventas de televisores, ya se tienen proyecciones desestacionalizadas para los cuatro trimestres siguientes. Ahora es necesario ajustar las proyecciones de acuerdo al efecto estacional. El índice estacional para el primer trimestre del año 5 ($t = 17$) es 0.93, de manera que el pronóstico para ese trimestre se obtiene al multiplicar el pronóstico desestacionalizado basado en la tendencia ($T_{17} = 7\,617$) por el índice estacional (0.93). Por tanto, el pronóstico para el trimestre siguiente es 7 617(0.93) = 7 084. En la tabla 18.12 se presentan los pronósticos trimestrales para los trimestres 17 a 20. El cuarto trimestre de alto volumen de ventas tiene un pronóstico de 9 190 unidades, y el segundo trimestre de volumen bajo de ventas tiene 6 523 unidades como pronóstico.

Modelos basados en datos mensuales

En el ejemplo de las ventas de televisores se emplearon datos trimestrales para ilustrar el cálculo de los índices estacionales. Sin embargo, en muchas ocasiones suelen usarse pronósticos mensuales en lugar de pronósticos trimestrales. En tales casos el procedimiento presentado en esta sección puede emplearse con ligeros cambios. Primero, en lugar de un promedio móvil de cuatro trimestres se usa un promedio móvil de 12 meses; segundo, se calculan índices estacionales de 12 meses en lugar de índices estacionales de cuatro trimestres. Fuera de estos cambios, los cálculos y los pronósticos son idénticos.

Componente cíclico

En términos matemáticos es posible ampliar el modelo multiplicativo de la ecuación (18.9) para incluir el componente cíclico.

$$Y_t = T_t \times C_t \times S_t \times I_t \tag{18.10}$$

El componente cíclico, como ocurre con el componente estacional, se expresa como porcentaje de la tendencia. Como se dijo en la sección 18.1, este componente se atribuye a ciclos multianuales en la serie de tiempo. Es semejante al componente estacional, sólo que se presenta a lo largo de periodos más prolongados. Sin embargo, debido a la duración del componente cíclico, suele ser difícil obtener suficientes datos relevantes para estimarlo. Otra dificultad es que estos ciclos suelen tener longitudes variables. Un estudio más detallado del componente cíclico se deja para libros sobre métodos de pronóstico.

Ejercicios

Métodos



22. Considere los datos de la siguiente serie de tiempo.

Trimestre	Año 1	Año 2	Año 3
1	4	6	7
2	2	3	6
3	3	5	6
4	5	7	8

a. Dé los valores de los promedios móviles de cuatro trimestres y de los promedios móviles centrados.
b. Calcule los índices estacionales de los cuatro trimestres.

Aplicaciones

23. A continuación se presentan los datos, correspondientes a los últimos tres años de ventas trimestrales (número de ejemplares vendidos) de un libro de texto universitario.

Trimestre	Año 1	Año 2	Año 3
1	1690	1800	1850
2	940	900	1100
3	2625	2900	2930
4	2500	2360	2615

a. Para esta serie de tiempo dé los promedios móviles de cuatro trimestres y los promedios móviles centrados.
b. Calcule los índices estacionales de los cuatro trimestres.
c. ¿Cuándo obtiene la editorial el mayor índice estacional? ¿Parece ser razonable este resultado? Explique.

24. A continuación se presentan los gastos mensuales, a lo largo de tres años, en un edificio de seis departamentos en el sur de Florida. Determine los índices estacionales mensuales. Use 12 meses como promedio móvil.



	Gastos		
	Año 1	Año 2	Año 3
Enero	170	180	195
Febrero	180	205	210
Marzo	205	215	230
Abril	230	245	280
Mayo	240	265	290
Junio	315	330	390
Julio	360	400	420
Agosto	290	335	330
Septiembre	240	260	290
Octubre	240	270	295
Noviembre	230	255	280
Diciembre	195	220	250

25. En el sur de California, los especialistas en el control de la contaminación atmosférica cada hora monitorean las cantidades de ozono, dióxido de carbono y dióxido de nitrógeno en el aire. En los datos de esta serie de tiempo horaria se observa estacionalidad, los niveles de contaminación muestran ciertos patrones según la hora del día. Los niveles de dióxido de nitrógeno en el centro, para las 12 horas, de las 6:00 de la mañana a las 6:00 de la tarde, los días 15, 16 y 17 de julio fueron los siguientes.

archivo en CD
Pollution

15 de julio:	25	28	35	50	60	60	40	35	30	25	25	20
16 de julio:	28	30	35	48	60	65	50	40	35	25	20	20
17 de julio:	35	42	45	70	72	75	60	45	40	25	25	25

a. Determine los índices estacionales por hora de las 12 lecturas de cada día.
b. Mediante los índices estacionales del inciso a, se desestacionalizaron los datos; la ecuación de tendencia obtenida para los datos desestacionalizados es $T_t = 32.983 + 0.3922t$. Emplee únicamente el componente de tendencia y obtenga los pronósticos para las 12 horas del 18 de julio.
c. Use los índices de tendencia del inciso a para ajustar los pronósticos de tendencia obtenidos en el inciso b.

26. El consumo de energía eléctrica se mide en kilowatts-hora (kWh). La empresa pública local que proporciona el servicio de energía eléctrica ofrece un programa de ahorro en el que los clientes comerciales participantes pagan tarifas especialmente favorables a condición de que reduzcan su consumo de energía eléctrica cuando la empresa pública se los solicite. La empresa Timko Products redujo su consumo de energía eléctrica a partir del mediodía del jueves. Para evaluar el ahorro de energía, la empresa pública tiene que estimar el consumo normal de energía de Timko. El periodo de reducción de consumo de energía fue desde el medio día hasta las 8:00 de la noche. Los datos sobre el consumo de energía de esta empresa en las 72 horas anteriores son los siguientes.


Power

Lapso	Lunes	Martes	Miércoles	Jueves
0:00-4:00	—	19 281	31 209	27 330
4:00-8:00	—	33 195	37 014	32 715
8:00-12:00	—	99 516	119 968	152 465
0:00-24:00	124 299	123 666	156 033	
4:00-8:00	113 545	111 717	128 889	
8:00-24:00	41 300	48 112	73 923	

a. ¿Se observa algún efecto estacional en este periodo de 24 horas? Calcule los índices estacionales de estos seis lapsos de 4 horas.
b. Emplee el ajuste de tendencia para estimar los índices estacionales del consumo normal de Timko en el periodo que realizó el ahorro.



18.5 Análisis de regresión

Cuando se estudió el análisis de regresión en los capítulos 14, 15 y 16, se mostró cómo usar una o varias variables independientes para predecir el valor de una variable dependiente. Si considera el análisis de regresión como una herramienta para pronóstico, el valor de la serie de tiempo que se desea pronosticar puede verse como la variable dependiente. Por tanto, si se logra determinar un buen conjunto de variables independientes, o predictoras, se podrá obtener una ecuación de regresión estimada para predecir o pronosticar la serie de tiempo.

El método empleado en la sección 18.3 para ajustar una línea de tendencia lineal a la serie de tiempo de las ventas de bicicletas es un caso especial del análisis de regresión. En ese ejemplo se mostró que las dos variables, ventas de bicicletas y tiempo, estaban relacionadas linealmente.* Debido a la inherente complejidad de la mayoría de los problemas reales, para predecir

*En un sentido estrictamente técnico, no se considera que el número de bicicletas vendidas esté relacionado con el tiempo, sino que el tiempo se usa como sustituto de variables con las que está relacionada el número de bicicletas vendidas, pero tales variables no se conocen o son difíciles de medir.

la variable de interés es necesario considerar más de una variable. En tales situaciones se usa la técnica estadística conocida como análisis de regresión múltiple.

Recuerde que para obtener una ecuación estimada de regresión múltiple se necesita una muestra de observaciones de la variable dependiente y de todas las variables independientes. En el análisis de las series de tiempo, los datos de n periodos de la serie de tiempo representan una muestra de n observaciones de cada una de las variables que pueden usarse en el análisis. La notación que se usa para una función con k variables independientes es la siguiente.

$$\begin{aligned} Y_t &= \text{valor de la serie de tiempo en el periodo } t \\ x_{1t} &= \text{valor de la variable independiente 1 en el periodo } t \\ x_{2t} &= \text{valor de la variable independiente 2 en el periodo } t \\ &\;\;\vdots \\ x_{kt} &= \text{valor de la variable independiente } k \text{ en el periodo } t \end{aligned}$$

Los n periodos de datos que se necesitan para obtener la ecuación estimada de regresión se verán como se muestra en la tabla siguiente.

Periodo	Serie de tiempo (Y_t)	Valor de las variables independientes						
		x_{1t}	x_{2t}	x_{3t}	.	.	.	x_{kt}
1	Y_1	x_{11}	x_{21}	x_{31}	.	.	.	x_{k1}
2	Y_2	x_{12}	x_{22}	x_{32}	.	.	.	x_{k2}
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
n	Y_n	x_{1n}	x_{2n}	x_{3n}	.	.	.	x_{kn}

Como puede uno suponer, existen diversas posibilidades para elegir las variables independientes para el modelo de pronóstico. Una posible variable independiente es simplemente el tiempo. Ésta fue la variable que se eligió en la sección 18.3 cuando se estimó la tendencia de la serie de tiempo mediante una función lineal de la variable independiente tiempo. Si $x_{1t} = t$, se obtiene una ecuación de regresión estimada de la forma

$$\hat{Y}_t = b_0 + b_1 t$$

donde $\hat{Y}_t$ es una estimación del valor Y_t de la serie de tiempo y donde b_0 y b_1 son estimaciones de los coeficientes de regresión. En modelos más complejos, pueden tenerse más términos que corresponden al tiempo elevado a otras potencias. Por ejemplo, si $x_{2t} = t^2$ y $x_{3t} = t^3$, la ecuación estimada de regresión tomará la forma

$$\begin{aligned} \hat{Y}_t &= b_0 + b_1 x_{1t} + b_2 x_{2t} + b_3 x_{3t} \\ &= b_0 + b_1 t + b_2 t^2 + b_3 t^3 \end{aligned}$$

Observe que con este modelo se obtiene un pronóstico para una serie de tiempo que tiene características curvilíneas a lo largo del tiempo.

En otros modelos de pronóstico basados en la regresión se emplea una mezcla de variables independientes económicas y demográficas. Por ejemplo, para pronosticar las ventas de refrigeradores, pueden emplearse las siguientes variables independientes.

x_{1t} = precio en el periodo t
x_{2t} = ventas totales de la industria en el periodo $t - 1$
x_{3t} = número de permisos de construcción para casas en el periodo $t - 1$
x_{4t} = pronóstico poblacional para el periodo t
x_{5t} = presupuesto para publicidad en el periodo t

De acuerdo con el procedimiento usual de regresión múltiple, para obtener el pronóstico habrá que usar una ecuación estimada de regresión con cinco variables independientes.

Spyros Makridakis, un destacado experto en pronósticos, mostró que las técnicas más sencillas son mejores que los procedimientos más complicados para pronósticos a corto plazo. El uso de procedimientos más caros y sofisticados no garantiza la obtención de un mejor pronóstico.

La obtención de un buen pronóstico mediante un método de regresión depende en gran parte de la capacidad para identificar y obtener los datos de variables independientes que estén estrechamente relacionadas con la serie de tiempo. Por lo general, al tratar de obtener una ecuación estimada de regresión se ponen a consideración diversos conjuntos de variables independientes. Por tanto, una parte del proceso del análisis de regresión debe ser la elección del conjunto de variables independientes con las que se obtenga el mejor modelo de pronóstico.

En la introducción de este capítulo se dijo que en los **modelos causales de pronóstico** se usan series de tiempo relacionadas con la serie que se quiere pronosticar para tratar de explicar la causa del comportamiento de dicha serie. La herramienta más usada para obtener modelos causales es el análisis de regresión. La serie de tiempo relacionada es la variable independiente y la serie de tiempo que se quiere pronosticar es la variable dependiente.

En otro tipo de modelo de pronóstico basado en la regresión, las variables independientes son todos los valores anteriores de la misma serie de tiempo. Por ejemplo, si los valores de la serie de tiempo se denotan $Y_1, Y_2, ..., Y_n$, y la variable independiente es Y_t, se trata de hallar una ecuación de regresión estimada que relacione Y_t con los valores más recientes de la serie de tiempo Y_{t-1}, Y_{t-2}, etc. Si se emplean como variables independientes los tres periodos más recientes, la ecuación estimada de regresión será

$$\hat{Y}_t = b_0 + b_1 Y_{t-1} + b_2 Y_{t-2} + b_3 Y_{t-3}$$

A los modelos de regresión que tienen variables independientes con los valores anteriores de la serie de tiempo se les conoce como **modelos autorregresivos**.

Por último, en otro método de pronóstico basado en regresión se incorpora una mezcla de las variables independientes ya estudiadas. Por ejemplo, se puede usar una combinación de variables de tiempo, algunas variables económico-demográficas y algunos de los valores previos de las mismas series de tiempo.



Métodos cualitativos

Si no se cuenta con datos históricos, es necesario emplear técnicas cualitativas para obtener pronósticos. Pero el costo de emplear las técnicas cualitativas puede ser elevado por la cantidad de tiempo que se requiere invertir.

En las secciones anteriores se vieron varios métodos cuantitativos para hacer pronósticos. En la mayor parte de estas técnicas se necesitan datos históricos sobre la variable de interés, de manera que estas técnicas no se pueden emplear cuando no se cuenta con datos históricos. Además de esto, aun cuando se cuente con datos históricos, un cambio significativo que afecte a la serie de tiempo puede hacer cuestionable el uso de datos del pasado para predecir valores futuros de la serie de tiempo. Por ejemplo, un programa de racionalización de la gasolina, impuesto por el gobierno, hará dudar de la validez de un pronóstico sobre las ventas de gasolina que se base en datos históricos. Las técnicas cualitativas de pronóstico ofrecen una alternativa en éstas y otras situaciones.

Método de Delphi

Una de las técnicas cualitativas de pronóstico más usadas es el **método de Delphi**, elaborado por un grupo de investigadores de Rand Corporation. En este método se trata de obtener un pronóstico mediante un "consenso de grupo". En su modo usual de aplicación, se le pide a un panel de

expertos —que no se conocen entre sí y que se encuentran separados unos de otros— que respondan una serie de cuestionarios. Las respuestas del primer cuestionario se tabulan y se usan para elaborar un segundo cuestionario que contiene información y opiniones de todo el grupo. Después se le pide a cada uno de los participantes que reconsidere, y si es necesario, modifique su respuesta anterior a la luz de la información del grupo. Este proceso continúa hasta que el coordinador considere que se ha alcanzado cierto grado de consenso. El objetivo del método de Delphi no es dar como resultado una sola respuesta, sino una gama reducida de opiniones en las que coincidan la mayor parte de los expertos.

Opinión de un experto

Las evidencias empíricas y los argumentos teóricos indican que en pronósticos obtenidos mediante la opinión de un experto deben intervenir entre 5 y 20 expertos. Sin embargo, en situaciones en las que interviene el crecimiento exponencial, los pronósticos obtenidos mediante la opinión de un experto pueden no ser apropiados.

Los pronósticos cualitativos suelen estar basados en la opinión de un solo experto o representar el consenso de un grupo de expertos. Por ejemplo, cada año un grupo de expertos de Merril Lynch se reúnen para pronosticar el promedio industrial Dow Jones y su tipo de interés preferencial para el año siguiente. Para esto, cada uno de los expertos analiza información que considera con influencia sobre el mercado de acciones y sobre las tasas de interés; después combinan sus informaciones en un pronóstico. No se emplea ningún modelo formal y es poco probable que dos expertos analicen la misma información de la misma manera.

La opinión de los expertos es un método de pronóstico que suele recomendarse cuando es poco probable que las condiciones del pasado se presenten en el futuro. Aun cuando no se usa ningún modelo cuantitativo formal, este método ha dado buenos pronósticos en muchas situaciones.

Escenarios futuros

El método cualitativo conocido como **escenarios futuros** consiste en elaborar un escenario conceptual del futuro con base en un conjunto bien definido de suposiciones. Distintos conjuntos de suposiciones llevan a diferentes escenarios. La persona que debe tomar las decisiones tiene que decidir cuán probable es cada escenario y tomar las decisiones de acuerdo con ese escenario.

Métodos intuitivos

Los métodos *subjetivos* o *cualitativos intuitivos* se basan en que la mente humana tiene la capacidad de procesar una gran cantidad de información que, en la mayoría de los casos, sería difícil de cuantificar. Estas técnicas suelen usarse en trabajos de grupo, en donde un comité o panel trata de desarrollar ideas nuevas o de resolver problemas complejos a través de una "sesión de lluvia de ideas". En estas sesiones las personas son liberadas de las usuales restricciones o presiones de grupo y de las críticas, ya que pueden exponer cualquier idea u opinión sin importar su relevancia y, lo que es más importante, sin miedo a la crítica.

Resumen

En este capítulo se presentó una introducción a los métodos básicos de análisis de series de tiempo y de pronóstico. Primero se indicó que para explicar el comportamiento de una serie de tiempo, es útil entenderla como formada por cuatro componentes: un componente de tendencia, un componente estacional, un componente cíclico y un componente irregular. Al aislar estos componentes y medir su efecto, es posible pronosticar valores futuros de la serie de tiempo.

Se vio cómo emplear los métodos de suavizamiento para pronosticar una serie de tiempo que no presente efectos significativos de tendencia, estacionales o cíclicos. El método de los promedios móviles consiste en calcular un promedio de los valores de los datos del pasado y después usar ese promedio como pronóstico para el periodo siguiente. El método de suavizamiento exponencial, usa un promedio ponderado de los valores pasados de la serie de tiempo para calcular un pronóstico.

Se mostró cómo usar el análisis de regresión para hacer proyecciones de tendencia cuando la serie de tiempo únicamente muestra una tendencia a largo plazo. Cuando una serie de tiempo tiene tanto una influencia de tendencia como una influencia estacional significativas, se mostró cómo aislar los efectos de estos dos factores para obtener mejores pronósticos. Por último, se

describió el análisis de regresión como un procedimiento para obtener modelos causales de pronóstico. Un modelo causal de pronóstico es un modelo que relaciona los valores de la serie de tiempo (variable dependiente) con otras variables independientes que se cree explican (causan) el comportamiento de la serie de tiempo.

Los modelos cualitativos de pronóstico se trataron como modelos útiles cuando no se cuenta con datos históricos o cuando se cuenta con pocos datos históricos. Estos métodos también se usan cuando se espera que el patrón pasado de la serie de tiempo no sea el mismo en el futuro.

Glosario

Serie de tiempo Conjunto de observaciones correspondientes a los valores de una variable, medidas en puntos sucesivos a lo largo del tiempo o durante periodos sucesivos de tiempo.

Pronóstico Es la predicción de los valores futuros de una serie de tiempo.

Tendencia Desplazamiento o movimiento de la serie de tiempo a largo plazo, observable a través de varios periodos.

Componente cíclico El componente de una serie de tiempo que hace que ésta muestre un comportamiento que consiste en tendencias periódicas de aumento y disminución, tendencias que tienen una duración de más de un año.

Componente estacional El componente de una serie de tiempo que muestra que en ella existe un patrón periódico que dura un año o menos.

Componente irregular El componente de una serie de tiempo que corresponde a las variaciones aleatorias que se observan en los valores de la misma, variaciones que no son explicadas por los componentes de tendencia, cíclicos o estacionales.

Promedios móviles Método para obtener pronósticos o para suavizar una serie de tiempo, en el que como pronóstico para cada periodo siguiente se usa el promedio de los valores de los n datos más recientes de la serie de tiempo.

Cuadrado medio debido al error (CME) Es una medida de la exactitud que se obtiene con un método de pronóstico. Esta medida es el promedio de la suma de los cuadrados de las diferencias entre los valores pronosticados para la serie de tiempo y sus valores reales.

Promedios móviles ponderados Método que se emplea para obtener pronósticos o para suavizar una serie de tiempo mediante un promedio ponderado de los valores de datos pasados. La suma de los pesos empleados debe ser uno.

Suavizamiento exponencial Técnica de pronóstico en la que se emplea un promedio ponderado de valores pasados de la serie de tiempo.

Constante de suavizamiento Es el parámetro que se emplea en el modelo de suavizamiento exponencial como peso para el valor más reciente de la serie de tiempo.

Modelo multiplicativo para series de tiempo Modelo en el que se multiplican los diversos componentes de una serie de tiempo para obtener así el valor real de la serie de tiempo. Cuando los cuatro componentes, de tendencia, cíclico, estacional e irregular están presentes, se obtiene $Y_t = T_t \times C_t \times S_t \times I_t$. Cuando el componente cíclico no está modelado se obtiene $Y_t = T_t \times S_t \times I_t$.

Serie de tiempo desestacionalizada Serie de tiempo de la que se ha eliminado el efecto estacional. Esto se hace al dividir cada observación de la serie de tiempo original entre su correspondiente índice estacional.

Métodos causales de pronóstico Métodos para obtener pronósticos en los que una serie de tiempo se relaciona con otras variables que se considera explican o causan el comportamiento de la serie de tiempo.

Modelo autorregresivo Modelo para predecir valores futuros de una serie de tiempo en el que se usa una relación de regresión con base en valores pasados de la serie de tiempo.

Método de Delphi Método cualitativo de pronóstico en el que los pronósticos se obtienen mediante consensos de grupo.

Escenarios futuros Método cualitativo de pronóstico que consiste en desarrollar un escenario conceptual futuro a partir de un conjunto bien definido de suposiciones.

Fórmulas clave

Promedio móvil

$$\text{Promedio móvil} = \frac{\Sigma(\text{de los valores de los } n \text{ datos más recientes})}{n} \quad (18.1)$$

Modelo de suavizamiento exponencial

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t \quad (18.2)$$

Ecuación de tendencia lineal

$$T_t = b_0 + b_1 t \quad (18.5)$$

Modelo multiplicativo para series de tiempo con los componentes de tendencia, estacional e irregular

$$Y_t = T_t \times S_t \times I_t \quad (18.9)$$

Modelo multiplicativo para series de tiempo con los componentes de tendencia, cíclico, estacional e irregular

$$Y_t = T_t \times C_t \times S_t \times I_t \quad (18.10)$$

Ejercicios complementarios

27. Los promedios móviles suelen usarse para identificar movimientos en los precios de las acciones (en dólares por acción). A continuación se presentan los precios de cierre de IBM desde el 24 de agosto de 2004 hasta 16 de agosto de 2005 (*Compustat*, 26 de febrero de 2006).



IBM

Día	Precio ($)	Día	Precio ($)
24 de agosto	81.32	7 de septiembre	80.98
25 de agosto	81.10	8 de septiembre	80.80
26 de agosto	80.38	9 de septiembre	81.44
29 de agosto	81.34	12 de septiembre	81.48
30 de agosto	80.54	13 de septiembre	80.75
31 de agosto	80.62	14 de septiembre	80.48
1 de septiembre	79.54	15 de septiembre	80.01
2 de septiembre	79.46	16 de septiembre	80.33
6 de septiembre	81.02		

a. Use un promedio móvil de tres días para suavizar la serie de tiempo. Pronostique el precio de cierre del 19 de septiembre del 2005 (que es el siguiente día de operaciones).
b. Emplee el suavizamiento exponencial con $\alpha = 0.6$ como constante de suavizamiento para suavizar la serie de tiempo. Pronostique el precio de cierre del 19 de septiembre del 2005 (que es el siguiente día de operaciones).
c. ¿Cuál de los dos métodos prefiere? ¿Por qué?

28. En el 2005 los ingresos que obtuvo Xerox Corporation por sus productos y servicios de color fueron de $4.6 mil millones, 30% del total de sus ingresos. En la tabla siguiente se presentan las variaciones porcentuales trimestrales a lo largo de 12 trimestres (*Democrat and Chronicle*, 5 de marzo de 2006).

Año	Trimestre	Crecimiento (%)
2003	1	15
	2	19
	3	15
	4	20
2004	1	26
	2	17
	3	18
	4	21
2005	1	15
	2	17
	3	22
	4	17

a. Use el suavizamiento exponencial para pronosticar la serie de tiempo. Emplee las constantes de suavizamiento $\alpha = 0.1$, $\alpha = 0.2$, $\alpha = 0.3$.
b. ¿Con cuál de estos valores de suavizamiento se obtiene un mejor pronóstico?

29. En la tabla siguiente se presentan los porcentajes de acciones en un portafolio estándar a lo largo de nueve trimestres que van desde el 2005 hasta el 2007.

Trimestre	Acciones (%)
1-2005	29.8
2-2005	31.0
3-2005	29.9
4-2005	30.1
1-2006	32.2
2-2006	31.5
3-2006	32.0
4-2006	31.9
1-2007	30.0

a. Emplee el suavizamiento exponencial para esta serie de tiempo. Utilice las constantes de suavizamiento $\alpha = 0.2$, $\alpha = 0.3$ y $\alpha = 0.4$. ¿Con cuál de estos valores de la constante de suavizamiento se obtiene un mejor pronóstico?
b. Pronostique, para el segundo trimestre de 2007, el porcentaje de acciones en un portafolio estándar.

30. Una cadena de tiendas de abarrotes registró la demanda semanal (en cajas) de un determinado detergente para trastos. Estos datos se presentan en la tabla siguiente. Emplee el suavizamiento exponencial con $\alpha = 0.2$ y obtenga un pronóstico para la semana 11.

Semana	Demanda	Semana	Demanda
1	22	6	24
2	18	7	20
3	23	8	19
4	21	9	18
5	17	10	21

31. United Dairies, Inc., es el proveedor de leche de varias empresas de abarrotes en el condado Dade de Florida. Los gerentes de United Dairies desean contar con un pronóstico del número de medios galones de leche que se venden por semana. A continuación se presentan los datos de las ventas en las últimas 12 semanas.

Semana	Ventas	Semana	Ventas
1	2750	7	3300
2	3100	8	3100
3	3250	9	2950
4	2800	10	3000
5	2900	11	3200
6	3050	12	3150

Use el suavizamiento exponencial con $\alpha = 0.4$ para obtener un pronóstico de demanda para la semana 13.

32. El grupo Garden Avenue Seven vende discos compactos de sus presentaciones. En la tabla siguiente se presentan las ventas (en unidades) en los últimos 18 meses. El administrador del grupo desea contar con un método exacto para pronosticar las ventas.



CDSales

Mes	Ventas	Mes	Ventas	Mes	Ventas
1	293	7	381	13	549
2	283	8	431	14	544
3	322	9	424	15	601
4	355	10	433	16	587
5	346	11	470	17	644
6	379	12	481	18	660

a. Emplee el suavizamiento exponencial con $\alpha = 0.3$, 0.4, y 0.5. ¿Con cuál de estos valores de α obtiene mejores pronósticos?
b. Haga un pronóstico mediante la proyección de tendencia. Dé el valor del CME.
c. ¿Qué método de pronóstico le recomendaría usted al administrador? ¿Por qué?

33. La tienda departamental Mayfair, que se encuentra en Davenport, Iowa (Estados Unidos), necesita determinar la pérdida de ventas que tuvo durante los meses de julio y agosto, en los que tuvo que cerrar a causa de los daños sufridos por el desbordamiento del río Mississippi. A continuación se presentan los datos de las ventas desde enero hasta junio.

Mes	Ventas ($ miles)	Mes	Ventas ($ miles)
Enero	185.72	Abril	210.36
Febrero	167.84	Mayo	255.57
Marzo	205.11	Junio	261.19

a. Emplee el suavizamiento exponencial con $\alpha = 0.4$ y obtenga un pronóstico para julio y agosto. (*Sugerencia:* para obtener el pronóstico para agosto, emplee el pronóstico para julio como ventas reales de julio.) Dé un comentario sobre el uso del suavizamiento exponencial para pronosticar más de un periodo futuro.
b. Use la proyección de tendencia para pronosticar las ventas en julio y agosto.
c. La aseguradora de Mayfair propuso una liquidación de $240 000 por la pérdida de las ventas de julio y agosto. ¿Es una cantidad justa? Si no es así, ¿qué cantidad recomendaría usted como contraoferta?

34. Canton Supplies, Inc., es una empresa de servicios que emplea a 100 individuos, aproximadamente. A los gerentes de la empresa les preocupa el cumplimiento de sus obligaciones en efecti-

vo por lo que desean obtener un pronóstico de los requerimientos mensuales de efectivo. Debido a un cambio reciente en la política de operación, únicamente se consideran relevantes los últimos siete meses. A partir de la proyección de tendencia y los datos históricos siguientes, pronostique los requerimientos de efectivo en los dos próximos meses.

Mes	1	2	3	4	5	6	7
Efectivo requerido ($ miles)	205	212	218	224	230	240	246

35. A continuación se presentan los saldos mínimos promedio en cuentas de cheques que pagan intereses para evitar tener que pagar cargos; éstos fueron los saldos mínimos vigentes desde el año 2000 hasta el año 2006 (*USA Today*, 6 de diciembre de 2005).



AcctBal

Fecha	**Saldo ($)**
Primavera 2000	1 522.41
Otoño 2000	1 659.63
Primavera 2001	1 678.34
Otoño 2001	1 707.55
Primavera 2002	1 767.36
Otoño 2002	1 866.17
Primavera 2003	2 015.04
Otoño 2003	2 257.82
Primavera 2004	2 425.83
Otoño 2004	2 086.93
Primavera 2005	2 295.85
Otoño 2005	2 294.61

a. Grafique esta serie de tiempo. ¿Parece haber una tendencia lineal?
b. Obtenga la ecuación de tendencia lineal para esta serie de tiempo.
c. Utilice la ecuación de tendencia para pronosticar el saldo promedio mínimo para evitar pagar recargos en la primavera de 2006.

36. La empresa Costello Music tiene cinco años de existencia. En este lapso las ventas de pianos aumentaron de 12 pianos en el primer año a 76 pianos en el último año. Fred Costello, el dueño de la empresa, desea pronosticar la venta de pianos del año próximo. A continuación se presentan los datos históricos.

Año	1	2	3	4	5
Ventas	12	28	34	50	76

a. Grafique esta serie de tiempo. ¿Parece seguir una tendencia lineal?
b. Obtenga la ecuación para el componente de tendencia de esta serie de tiempo. ¿Cuál es el crecimiento anual promedio que ha tendido la empresa?

37. Durante los últimos siete años, la empresa Hudson Marine ha sido distribuidor autorizado de los radios náuticos de C&D. En la tabla siguiente se da el número de radios vendidos por año por esa empresa.

Año	1	2	3	4	5	6	7
Número vendido	35	50	75	90	105	110	130

a. Trace la gráfica de esta serie de tiempo.
b. Obtenga la ecuación de tendencia lineal de esta serie de tiempo.
c. A partir de la ecuación de tendencia lineal obtenida en el inciso b pronostique las ventas anuales del año 8.

38. La League of American Theatres and Producers, Inc., recaba diversos datos estadísticos sobre los espectáculos que se presentan en Broadway, como ingreso bruto, tiempo que se mantiene el es-

pectáculo en escena y número de producciones nuevas. En la tabla siguiente se presenta la audiencia, por temporada (en millones), en los espectáculos de Broadway desde 1990 hasta 2001 (*The World Almanac*, 2002).

Temporada	Audiencia (en millones)	Temporada	Audiencia (en millones)
1990-1991	7.3	1996-1997	10.6
1991-1992	7.4	1997-1998	11.5
1992-1993	7.9	1998-1999	11.7
1993-1994	8.1	1999-2000	11.4
1994-1995	9.0	2000-2001	11.9
1995-1996	9.5		

a. Trace la gráfica de esta serie de tiempo y diga si es adecuado considerar que hay una tendencia lineal.
b. Dé la ecuación para el componente de tendencia lineal de esta serie de tiempo.
c. En esta serie de tiempo, ¿cuál es el incremento promedio, por temporada, que hay en la audiencia?
d. Emplee la ecuación de tendencia para pronosticar la audiencia en la temporada 2001- 2002.

39. En los últimos 25 años, la United States Golg Association (USGA) ha probado miles de pelotas de golf para ver si satisfacen los requerimientos de distancia. En la tabla siguiente se presenta el número de pelotas de golf probadas anualmente por la USGA desde 1992 hasta 2002 (*Golf Journal*, octubre de 2002).

Año	Número	Año	Número
1992	465	1997	919
1993	602	1998	916
1994	646	1999	861
1995	755	2000	834
1996	807	2001	821

Grafique esta serie de tiempo y haga un comentario si observa una tendencia lineal. ¿Qué tipo de función cree usted que sería la más adecuada para el patrón de tendencia que se observa en esta serie?

40. Regrese al ejercicio 37 sobre la empresa Hudson Marine. Suponga que las ventas trimestrales en los siete años de datos históricos son las siguientes.

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total de ventas anuales
1	6	15	10	4	35
2	10	18	15	7	50
3	14	26	23	12	75
4	19	28	25	18	90
5	22	34	28	21	105
6	24	36	30	20	110
7	28	40	35	27	130

a. Para esta serie dé los promedios móviles de cuatro trimestres. En una misma gráfica, trace tanto la serie de tiempo original como la serie de promedios móviles.
b. Calcule el índice estacional de los cuatro trimestres.
c. ¿Cuándo la empresa Hudson Marine experimenta el mayor efecto estacional? ¿Es razonable? Explique.

41. Vuelva al ejercicio 36 que trata de la empresa Costello Music. A continuación se presentan los datos de las ventas trimestrales.

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Ventas anuales totales
1	4	2	1	5	12
2	6	4	4	14	28
3	10	3	5	16	34
4	12	9	7	22	50
5	18	10	13	35	76

a. Calcule los índices estacionales de los cuatro trimestres.
b. ¿Cuándo experimenta el mayor efecto estacional? ¿Es razonable? Explique.

42. Vuelva a los datos de la empresa Hudson Marine presentados en el ejercicio 40.
a. Desestacionalice los datos y emplee la serie de tiempo desestacionalizada para determinar la tendencia.
b. Emplee los resultados del inciso a para obtener un pronóstico trimestral para el año próximo a partir de la tendencia.
c. Emplee los índices estacionales obtenidos en el ejercicio 40 para ajustar los pronósticos obtenidos en el inciso b de acuerdo con los efectos estacionales.

43. Regrese al ejercicio 41 sobre la empresa Costello Music.
a. Desestacionalice los datos y emplee la serie de tiempo desestacionalizada para determinar la tendencia.
b. Con los resultados del inciso a obtenga un pronóstico trimestral para el año próximo con base en la tendencia.
c. Emplee los índices estacionales obtenidos en el ejercicio 41 para ajustar los pronósticos obtenidos en el inciso b de acuerdo con los efectos estacionales.

Caso problema 1 Pronóstico para las ventas de alimentos y bebidas

El restaurante Vintage, en la Isla Captiva, cerca de Fort Myers, Florida, es operado por su dueña Karen Payne. Este restaurante acaba de cumplir tres años de funcionamiento. Durante este tiempo, Karen ha tratado de que el restaurante se dé a conocer como un establecimiento de alta calidad, especializado en mariscos frescos. Gracias al esfuerzo de Karen y su equipo, este restaurante se ha convertido en uno de los restaurantes mejores y de mayor crecimiento de la isla.

Karen considera que para planear el crecimiento del restaurante en el futuro, necesita elaborar un sistema que le permita pronosticar las ventas mensuales de alimentos y bebidas con hasta un año de anticipación. Karen ha reunido los datos siguientes sobre las ventas totales de alimentos y bebidas (dados en miles de dólares) durante estos tres años de funcionamiento.

archivo en CD
Vintage

Mes	Primer año	Segundo año	Tercer año
Enero	242	263	282
Febrero	235	238	255
Marzo	232	247	265
Abril	178	193	205
Mayo	184	193	210
Junio	140	149	160
Julio	145	157	166
Agosto	152	161	174
Septiembre	110	122	126
Octubre	130	130	148
Noviembre	152	167	173
Diciembre	206	230	235

Informe administrativo

Analice estos datos del restaurante Vintage. Redacte un informe para Karen en el que resuma sus hallazgos, pronósticos y sugerencias. El informe debe contener:

1. Una gráfica de la serie de tiempo.
2. El análisis de la estacionalidad de los datos. Indique el índice estacional para cada mes y comente sobre las ventas mensuales en las estaciones baja y alta.
3. Un pronóstico de ventas, de enero a diciembre, para el cuarto año.
4. Una sugerencia de cuándo deben actualizarse los datos de manera que se tomen en cuenta los nuevos datos de las ventas.
5. En un apéndice muestre todos los detalles de los cálculos.

Suponga que en enero del cuarto año las ventas resultan ser de $295 000. ¿De cuánto fue su error de pronóstico? Si este error es grande, a Karen le desconcertará esta diferencia entre su pronóstico y las ventas reales. ¿Qué puede hacer para despejar sus dudas sobre el procedimiento de pronóstico?

Caso problema 2 Pronóstico de pérdidas de ventas

La tienda de departamentos Carlson sufrió severos daños con la llegada del huracán del 31 de agosto de 2006. La tienda tuvo que permanecer cerrada cuatro meses (desde septiembre de 2006 hasta diciembre de 2006) y ahora se encuentra inmersa en una disputa, con su aseguradora, sobre el monto de las ventas perdidas durante el tiempo que permaneció cerrada. Los dos puntos más importantes a resolver son: 1) el monto de las ventas que hubiera hecho Carlson si no hubiera ocurrido el huracán y 2) si Carlson debe recibir una compensación por las ventas extra debidas al aumento de la actividad comercial después del huracán. El condado recibió más de 8 mil millones de dólares en ayuda federal para desastres y por el pago de seguros, lo que produjo un aumento de las ventas en tiendas departamentales y en muchos otros negocios.

En la tabla 18.13 se presentan los datos de las ventas de Carlson en los 48 meses antes del huracán. En la tabla 18.14 aparece el total de ventas en todas las tiendas departamentales en los 48 meses antes del huracán, así como el total de ventas en el condado en los cuatro meses en que estuvo cerrada la tienda departamental Carlson. Los gerentes de Carlson le piden a usted que analice estos datos y obtenga una estimación de la pérdida en ventas que sufrió Carlson en los cuatro meses que estuvo cerrada, de septiembre a diciembre de 2006. También le piden que determine

TABLA 18.13 VENTAS DE LA TIENDA DEPARTAMENTAL CARLSON DESDE SEPTIEMBRE DE 2002 HASTA AGOSTO DE 2006 ($ MILLONES)

Mes	2002	2003	2004	2005	2006
Enero		1.45	2.31	2.31	2.56
Febrero		1.80	1.89	1.99	2.28
Marzo		2.03	2.02	2.42	2.69
Abril		1.99	2.23	2.45	2.48
Mayo		2.32	2.39	2.57	2.73
Junio		2.20	2.14	2.42	2.37
Julio		2.13	2.27	2.40	2.31
Agosto		2.43	2.21	2.50	2.23
Septiembre	1.71	1.90	1.89	2.09	
Octubre	1.90	2.13	2.29	2.54	
Noviembre	2.74	2.56	2.83	2.97	
Diciembre	4.20	4.16	4.04	4.35	

TABLA 18.14 VENTAS EN LAS TIENDAS DEPARTAMENTALES DEL CONDADO, DESDE SEPTIEMBRE DE 2002 HASTA DICIEMBRE DE 2006 ($ MILLONES)

Mes	2002	2003	2004	2005	2006
Enero		46.8	46.8	43.8	48.0
Febrero		48.0	48.6	45.6	51.6
Marzo		60.0	59.4	57.6	57.6
Abril		57.6	58.2	53.4	58.2
Mayo		61.8	60.6	56.4	60.0
Junio		58.2	55.2	52.8	57.0
Julio		56.4	51.0	54.0	57.6
Agosto		63.0	58.8	60.6	61.8
Septiembre	55.8	57.6	49.8	47.4	69.0
Octubre	56.4	53.4	54.6	54.6	75.0
Noviembre	71.4	71.4	65.4	67.8	85.2
Diciembre	117.6	114.0	102.0	100.2	121.8

si se puede solicitar un pago por las ventas extras relacionadas con el huracán. Si se puede solicitar este pago, Carlson debe recibir una compensación por lo que hubiera ganado por las ventas extras además de sus ventas normales.

Informe administrativo

Redacte un informe para los directivos de la tienda departamental Carlson, en el que resuma sus hallazgos, sus pronósticos y sus sugerencias. El informe debe contener:

1. Una estimación de las ventas que se hubieran hecho de no haber habido huracán.
2. Una estimación de las ventas en las tiendas de departamentos de todo el condado si no se hubiera presentado el huracán.
3. Una estimación de la pérdida en ventas que sufrió la tienda departamental Carlson desde septiembre hasta diciembre de 2006.

Además, use las ventas reales en las tiendas departamentales de todo el condado, de septiembre a diciembre de 2006, y la estimación de la parte 2 para solicitar una indemnización por las ventas extra relacionadas con el huracán.

Apéndice 18.1 Pronósticos con Minitab

En este apéndice se muestra el uso de Minitab para hacer pronósticos con tres métodos de pronóstico: promedios móviles, suavizamiento exponencial y proyección de tendencia.

Promedios móviles



Gasoline

Para mostrar cómo usar Minitab para obtener pronósticos mediante el método de promedios móviles se emplearán los datos presentados en la tabla 18.1 y en la figura 18.5 de la serie de tiempo de las ventas de gasolina. Los datos de las ventas de gasolina en las 12 semanas se ingresan en la columna 2 de la hoja de cálculo. Para obtener un pronóstico para la semana 13, para promedios móviles de tres semanas, se siguen los pasos que se presentan a continuación.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Time Series**
Paso 3. Elegir **Moving Average**

Paso 4. Cuando aparezca el cuadro de diálogo Moving Average:
Ingresar C2 en el cuadro **Variable**
Ingresar 3 en el cuadro **MA length**
Seleccionar **Generate forecasts**
Ingresar 1 en el cuadro **Number of forecasts**
Ingresar 12 en el cuadro **Starting from origin**
Clic en **OK**

En la ventana de la sesión aparecerá el pronóstico para la semana 13 obtenido mediante promedios móviles de tres semanas. En los resultados de Minitab, el cuadrado medio debido al error, que es 10.22 aparece junto al rótulo MSD. Minitab cuenta con otras muchas opciones para dar los resultados, como dar una tabla resumen similar a la tabla 18.2 o una gráfica similar a la de la figura 18.6.

Suavizamiento exponencial



Para mostrar cómo usar Minitab para obtener pronósticos mediante el método de suavizamiento exponencial se emplearán nuevamente los datos presentados en la tabla 18.1 y en la figura 18.5 de la serie de tiempo de las ventas de gasolina. Los datos de las ventas de gasolina, en las 12 semanas, se ingresan en la columna 2 de la hoja de cálculo. Para obtener un pronóstico para la semana 13, usando como constante de suavizamiento $\alpha = 0.2$, se siguen los pasos que se presentan a continuación.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Time Series**
Paso 3. Elegir **Single Exp Smoothing**
Paso 4. Cuando aparezca el cuadro de diálogo Single Exp Smoothing:
Ingresar C2 en el cuadro **Variable**
Seleccionar la opción **Use** como Weight to Use in Smoothing
Ingresar 0.2 en el cuadro Use
Seleccionar **Generate forecasts**
Ingresar 1 en el cuadro **Number of forecasts**
Ingresar 12 en el cuadro **Starting from origin**
Seleccionar **Options**
Paso 5. Cuando aparezca el cuadro de diálogo Single Exp Smoothing – Options:
Ingresar 1 en el cuadro **Use average of first**
Clic en **OK**
Paso 6. Cuando aparezca el cuadro de diálogo Single Exp Smoothing:
Clic en **OK**

En la ventana de la sesión aparecerá el pronóstico para la semana 13 obtenido mediante suavizamiento exponencial. En los resultados de Minitab, el cuadrado medio del error aparece junto al rótulo MSD.* Minitab cuenta con otras muchas opciones para presentar los resultados, como dar una tabla resumen similar a la tabla 18.3 o una gráfica similar a la figura 18.7.

Proyección de tendencia



Para mostrar cómo usar Minitab para obtener pronósticos mediante la proyección de tendencia se emplearán los datos, presentados en la tabla 18.6 y en la figura 18.8, correspondientes a la serie de tiempo de las ventas de bicicletas. En la columna C1 se ingresa el número de años y en C2 los datos de las ventas. Para obtener un pronóstico para el año 11, emplee la proyección de tendencia, se siguen los pasos que se presentan a continuación.

*El valor MSD que da Minitab no es el mismo que el valor CME que parece en la tabla 18.4. Minitab usa el 17 como pronóstico para la semana 1, así que, para calcular el valor de MSD usa los datos de los 12 periodos de tiempo. En cambio, en la sección 18.2 el valor del CME se calculó empleando únicamente los datos desde las semanas 2 hasta 12, debido a que no se contaba con un valor pasado con el cual obtener un pronóstico para la semana 1.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Time Series**
Paso 3. Elegir **Trend Analysis**
Paso 4. Cuando aparezca el cuadro de diálogo Trend Analysis:
Ingresar C2 en el cuadro **Variable**
Elegir **Linear** como Model Type
Seleccionar **Generate forecasts**
Ingresar 1 en el cuadro **Number of forecasts**
Ingresar 10 en el cuadro **Starting from origin**
Clic en **OK**

En la ventana de la sesión aparecerá la ecuación de tendencia lineal y el pronóstico para el periodo siguiente.

Apéndice 18.2 Pronósticos con Excel

En este apéndice se muestra el uso de Excel para hacer pronósticos empleando tres métodos de pronóstico: promedios móviles, suavizamiento exponencial y proyección de tendencia.

Promedios móviles



Gasoline

Para mostrar cómo usar Excel para obtener pronósticos mediante el método de promedios móviles se emplearán los datos presentados en la tabla 18.1 y en la figura 18.5 pertenecientes a la serie de tiempo de las ventas de gasolina. Los datos de las ventas de gasolina en las 12 semanas se ingresan en los renglones 2 a 13 de la columna B de la hoja de cálculo. Para obtener un promedio móvil de tres semanas, se siguen los pasos que se presentan a continuación.

Paso 1. Seleccionar el menú **Herramientas**
Paso 2. Elegir **Análisis de datos**
Paso 3. Elegir **Media móvil** en la lista Funciones para análisis
Clic en **Aceptar**
Paso 4. Cuando aparezca el cuadro de diálogo Media móvil:
Ingresar B2:B13 en el cuadro **Rango de entrada**
Ingresar 3 en el cuadro **Intervalo**
Ingresar C2 en el cuadro **Rango de salida**
Clic en **Aceptar**

En la columna B de la hoja de cálculo aparecerán los pronósticos obtenidos mediante promedios móviles de tres semanas. También se pueden obtener pronósticos para periodos de una longitud distinta ingresando otro valor en el cuadro **Intervalo**.

Suavizamiento exponencial



Gasoline

Para mostrar cómo usar Excel para obtener pronósticos mediante el método de suavizamiento exponencial se emplearán nuevamente los datos presentados en la tabla 18.1 y en la figura 18.5 de la serie de tiempo de las ventas de gasolina. Los datos de las ventas de gasolina en las 12 semanas se ingresan en los renglones 2 a 13 de la columna B de la hoja de cálculo. Para obtener un pronóstico con la constante de suavizamiento $\alpha = 0.2$, se siguen los pasos que se presentan a continuación.

Paso 1. Seleccionar el menú **Herramientas**
Paso 2. Elegir **Análisis de datos**
Paso 3. Elegir **Suavizamiento exponencial** en la lista Funciones para análisis
Clic en **Aceptar**
Paso 4. Cuando aparezca el cuadro de diálogo Suavizamiento exponencial:
Ingresar B2:B13 en el cuadro **Rango de entrada**
Ingresar 0.8 en el cuadro **Factor de suavizamiento**
Ingresar C2 en el cuadro **Rango de salida**
Clic en **Aceptar**

En la ventana de la sesión aparecerán los pronósticos obtenidos mediante suavizamiento exponencial. Observe que el valor ingresado como Factor de suavizamiento es $1 - \alpha$; para obtener pronósticos con otras constantes de suavizamiento es necesario ingresar un valor diferente para $1 - \alpha$ en el cuadro para el Factor de suavizamiento.

Proyección de tendencia



Bicycle

Para mostrar cómo usar Excel para la proyección de tendencia se emplearán los datos, presentados en la tabla 18.6 y en la figura 18.8, correspondientes a la serie de tiempo de las ventas de bicicletas. Los datos, con sus correspondientes rótulos en el renglón 1, se ingresan en los renglones 1 a 11 de las columnas A y B. Para obtener un pronóstico para el año 11 con la proyección de tendencia se siguen los pasos que se presentan a continuación.

Paso 1. Seleccionar cualquier celda vacía de la hoja de cálculo
Paso 2. Seleccionar el menú **Insertar**
Paso 3. Elegir **Función**
Paso 4. Cuando aparezca el cuadro de diálogo Pegar función
Elegir **Estadísticas** del cuadro Categoría de la función
Elegir **Pronóstico,** del cuadro Nombre de la función
Clic en **OK**
Paso 5. Cuando aparezca el cuadro de diálogo Pronóstico
Ingresar 11 en el cuadro **x**
Ingresar B2:B11 en el cuadro **Conocido_y**
Ingresar A2:A11 en el cuadro **Conocido_x**
Clic en **OK**

El pronóstico para el año 11, en este caso, 32.5, aparecerá en la celda elegida en el paso 1.



CAPÍTULO 19

Métodos no paramétricos

CONTENIDO



LA ESTADÍSTICA *en* LA PRÁCTICA

WEST SHELL REALTORS*
CINCINNATI, OHIO

La empresa West Shell Realtors fue fundada en 1958 y en ese entonces contaba con una oficina y un equipo de ventas formado por tres personas. En 1964 inició un programa de expansión a largo plazo durante el cual, casi anualmente, abrió nuevas oficinas. Con el tiempo, West Shell creció hasta convertirse en la mayor empresa inmobiliaria de Greater Cincinnati y ahora cuenta con oficinas en el suroeste de Ohio, en el sureste de Indiana y en el norte de Kentucky.

A las empresas de bienes raíces, como West Shell, el análisis estadístico les sirve para monitorear el curso de sus ventas. Cada mes se presenta un informe de cada una de las oficinas de West Shell, así como del total de la empresa. Resumen estadístico sobre la cantidad total de dólares en ventas, número de unidades vendidas y precio mediano de venta por unidad son esenciales para mantener informados, tanto a los gerentes de las distintas oficinas, como a los gerentes generales sobre el progreso y los problemas de la organización.

Además de los resúmenes mensuales sobre el curso de las operaciones, la empresa emplea diversas consideraciones estadísticas como guía para sus planes y estrategias. West Shell ha implementado una estrategia de expansión planeada. Cada vez que, debido a este plan de expansión, se quiere abrir una nueva oficina de ventas, la empresa tiene que decidir dónde abrir la nueva oficina. El tipo de datos usados para evaluar y comparar las distintas alternativas para la ubicación de una nueva oficina son los precios de venta de las casas, las tasas de facturación y los volúmenes de ventas pronosticados. Con el fin de identificar si existía alguna diferencia entre los patrones de ventas de estas dos áreas, West Shell se valió de métodos estadísticos no paramétricos.

En una ocasión West Shell tenía dos zonas posibles para abrir una nueva oficina; Clifton y Roselawn. Al comparar las dos zonas se tomaron en consideración diversos factores, entre ellos, los precios de venta de las casas.



Para mantener su competitividad, West Shell realiza análisis estadísticos de los precios de las casas. © Cortesía de Coldwell Banker West Shell.

A partir de muestras de 25 ventas en Clifton y 18 ventas en Roselawn se eligió la prueba de la suma de los rangos de Mann-Whitney Wilcoxon como la prueba estadística adecuada para las diferencias entre los patrones de ventas. Con un nivel de significancia de 0.05, la prueba de Mann-Whitney-Wilcoxon no permitió rechazar la hipótesis nula de que las dos poblaciones de precios de venta fueran idénticas. Por tanto, West Shell tuvo que buscar otros criterios, distintos a los precios de venta de las casas, para su proceso de selección de la ubicación de su nueva oficina.

En este capítulo se mostrará cómo aplicar pruebas estadísticas no paramétricas como la prueba de Mann-Whitney-Wilcoxon. También se discutirá la interpretación adecuada de dichas pruebas.

* Los autores agradecen a Rodney Fightmaster de West Shell Realtors por proporcionar este artículo para *La estadística en la práctica*.

Los métodos estadísticos hasta ahora presentados en este libro se conocen como *métodos paramétricos*. En este capítulo se presentan varios **métodos no paramétricos**. Estos métodos suelen ser aplicables en las situaciones en que los métodos paramétricos no lo son. Los métodos no paramétricos suelen requerir suposiciones menos restrictivas acerca del nivel de medición de los datos y menos suposiciones acerca de la forma de las distribuciones de probabilidad generadas por los datos muestrales.

Una de las consideraciones para determinar si lo apropiado es un método paramétrico o un método no paramétrico es la escala de medición empleada para generar los datos. Todos los datos son generados por una de las cuatro escalas de medición: nominal, ordinal, de intervalo o de razón. Por tanto, todos los análisis estadísticos se realizan con datos ya sea nominales, ordinales, de intervalo o de razón.

A continuación se definen y se proporcionan ejemplos de cada una de las cuatro escalas de medición.

1. *Escala nominal.* Una escala de medición es nominal si los datos son etiquetas o categorías que se usan para definir un atributo de un elemento. Los datos nominales pueden ser numéricos o no numéricos.

   Ejemplos. El mercado en el que cotiza una acción (NYSE, NASDAQ o AMEX) es un dato nominal no numérico. El número de seguro social de una persona es un dato nominal numérico.
2. *Escala ordinal*. Una escala de medición es ordinal si los datos pueden usarse para jerarquizar u ordenar las observaciones. Los datos ordinales pueden ser numéricos o no numéricos.

   Ejemplos. Las medidas pequeño, mediano y grande para dar el tamaño de un objeto son datos ordinales no numéricos. El lugar de los individuos en una clase 1, 2, 3, … son datos ordinales numéricos.
3. *Escala de intervalo.* Una escala de medición es de intervalo si los datos tienen las propiedades de los datos ordinales y los intervalos entre observaciones se expresan en términos de una unidad de medición fija. Los datos de intervalo tienen que ser numéricos.

   Ejemplos. Las mediciones de temperatura son datos de intervalo. Suponga que la temperatura en un lugar es de 21°C y en otro es de 4°C. Estos lugares se pueden jerarquizar de acuerdo con lo calurosos que son: el primero es más caliente que el segundo. La unidad fija de medición, 1°C, permite decir cuán más caliente es el primer lugar: 17°C.
4. *Escala de razón*. Una escala de medición es de razón si los datos tienen las propiedades de los datos de intervalo y el cociente (o razón) entre dos medidas tiene sentido. Los datos de razón tienen que ser numéricos.

   Ejemplos. Variables como la distancia, la altura, el peso y el tiempo se miden con una escala de razón. Las mediciones de temperatura no son datos de razón debido a que no existe un punto cero definido intrínsecamente. Por ejemplo, el punto de congelación del agua en la escala Fahrenheit es 32 grados y en la escala Celsius es 0 grados. Los cocientes entre datos de temperatura no tienen sentido. Por ejemplo, no tiene sentido decir que cuando la temperatura ambiente es de 20 grados hace el doble de calor que cuando es de 10 grados.

En el capítulo 1 se dijo que con las escalas nominal y ordinal se obtienen datos cualitativos. Con las escalas de intervalo y de razón se obtienen datos cuantitativos.

La mayor parte de los métodos estadísticos conocidos como métodos paramétricos requieren el uso de datos de las escalas de intervalo o de razón. Con estos niveles de medición, tienen sentido las operaciones aritméticas y medias, varianzas, desviaciones estándar, etc., pueden calcularse, interpretarse y usarse en el análisis. Con datos nominales y ordinales no es apropiado calcular medias, varianzas ni desviaciones estándar; por tanto, no pueden emplearse los métodos paramétricos. La única manera de analizar esos datos para obtener conclusiones estadísticas es emplear los métodos no paramétricos.

Si el nivel de medición de los datos es nominal u ordinal, calcular la media, la varianza y la desviación estándar no tiene sentido. Por tanto, con este tipo de datos, muchos de los procedimientos estadísticos discutidos previamente no pueden emplearse.

En general, para que un método estadístico se clasifique como método no paramétrico, debe satisfacer, por lo menos, una de las condiciones siguientes.*

1. Ser un método que pueda ser usado con datos nominales.
2. Ser un método que pueda ser usado con datos ordinales.
3. Ser un método que pueda ser usado con datos de intervalo o de razón cuando no sea posible hacer suposiciones acerca de la forma de la distribución de la población.

Si el nivel de medición de los datos es de intervalo o de razón y si las suposiciones necesarias sobre la distribución de probabilidad de la población son apropiadas, con los métodos paramétricos se obtienen procedimientos estadísticos más potentes y más refinados. En muchos de los casos en que se puede aplicar tanto un método paramétrico como un método no paramétrico, el método no paramétrico es casi tan bueno o casi tan potente como el método paramétrico. En los casos en que los datos son nominales y ordinales o en los casos en que las suposiciones requeridas por los métodos paramétricos son inapropiadas, sólo se cuenta con los métodos no pa-

*Véase W. J. Conover, *Practical Nonparametric Statistics*, 3. ed. (John Wiley & Sons, 1998).

ramétricos. Debido a que en los métodos no paramétricos se requieren mediciones de los datos menos restrictivas y menos suposiciones acerca de la distribución de la población, se considera que tienen una aplicación más general que los métodos paramétricos. Los métodos no paramétricos que se presentan en este capítulo son la prueba de los signos, la prueba de los rangos con signo de Wilcoxon, la prueba de Mann-Whitney-Wilcoxon, la prueba de Kruskal-Wallis y la correlación de los rangos de Spearman.

19.1 Prueba de los signos

En una aplicación de investigación de mercado de la **prueba de los signos** se usa una muestra de n clientes potenciales para que indiquen su preferencia por una de dos marcas de un producto, por ejemplo, de un café, de un detergente o de un refresco. Las n expresiones de preferencia son datos nominales, ya que el consumidor simplemente nombra una preferencia. Dados estos datos, el objetivo es determinar si existe diferencia en las preferencias entre los dos artículos que se comparan. Como se verá, la prueba de los signos es un procedimiento estadístico no paramétrico para responder esta pregunta.

Caso de muestras pequeñas

El caso de la muestra pequeña es siempre que $n \leq 20$. A continuación, mediante un estudio realizado para Sun Coast Farms, se ilustra el uso de la prueba de los signos para el caso de una muestra pequeña; Sun Coast produce un jugo de naranja comercializado bajo el nombre Citrus Valley. Un competidor de Sun Coast Farms produce también un jugo de naranja que comercializa bajo el nombre de Tropical Orange. En un estudio acerca de las preferencias de los consumidores respecto a estas dos marcas, a 12 individuos se les dieron muestras, sin marca, de cada uno de los productos. La marca que cada individuo probó primero fue seleccionada aleatoriamente. Después de probar los dos productos, se pidió a estas personas que indicaran su preferencia por una de las dos marcas. En este estudio, el objetivo es ver si hay una preferencia de los consumidores por uno de los dos productos. Sea p la proporción de la población de consumidores que prefiere Citrus Valley; las hipótesis que se quiere probar son las siguientes.

TABLA 19.1 PROBABILIDADES BINOMIALES CON $n = 12$, $p = 0.50$

Número de signos más	Probabilidad
0	0.0002
1	0.0029
2	0.0161
3	0.0537
4	0.1208
5	0.1934
6	0.2256
7	0.1934
8	0.1208
9	0.0537
10	0.0161
11	0.0029
12	0.0002

$$H_0\colon p = 0.50$$
$$H_a\colon p \neq 0.50$$

Si no se rechaza H_0, no se tendrán evidencias que indiquen la existencia de alguna diferencia en las preferencias de los consumidores por estas dos marcas de jugos de naranja. Sin embargo, si se rechaza H_0, se podrá concluir que las preferencias de los consumidores hacia estas marcas son diferentes. En ese caso, la marca seleccionada por el mayor número de consumidores se considerará que es la marca preferida.

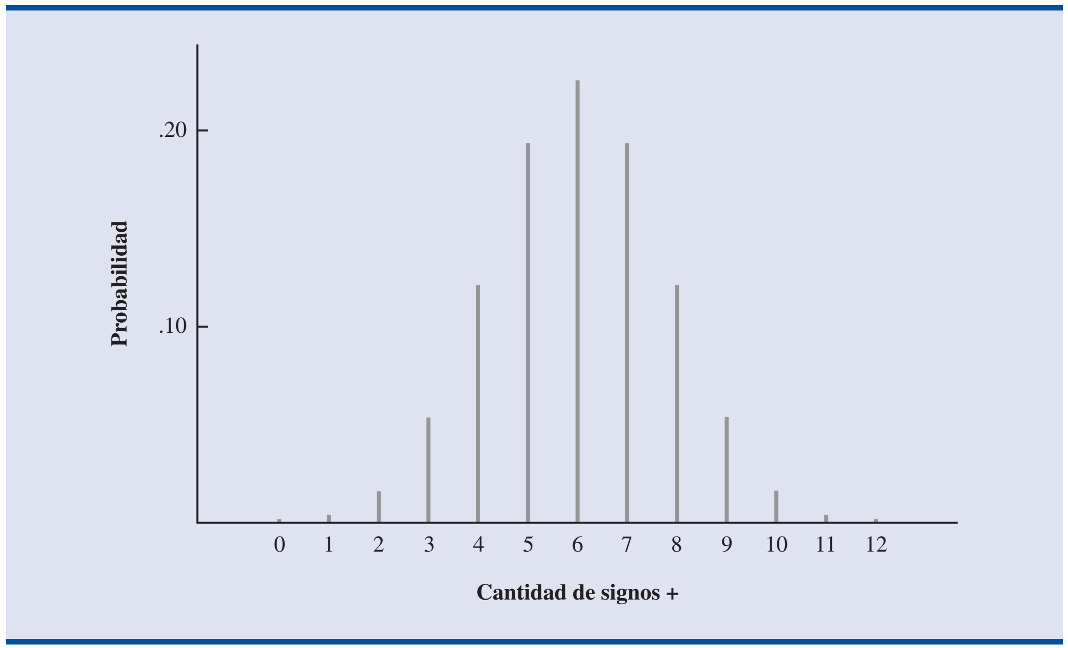
A continuación se muestra el uso de la versión para muestras pequeñas de la prueba de los signos al probar estas hipótesis para obtener una conclusión acerca de la preferencia de los consumidores. Para registrar los datos de la preferencia de los 12 individuos que participan en el estudio, se emplea un signo más si el individuo prefiere Citrus Valley y un signo menos si el individuo prefiere Tropical Orange. Debido a que los datos se registran en términos de signos más y menos, a esta prueba paramétrica se le conoce como prueba de los signos.

Las probabilidades binomiales exactas para tamaños de muestra menores o iguales a 20 se encuentran en la tabla 5 del apéndice B.

El número de signos más es el estadístico de prueba. Bajo la suposición de que H_0 es verdadera ($p = 0.50$), la distribución muestral del estadístico de prueba es una distribución binomial con $p = 0.50$. En la tabla 5 del apéndice B se encuentran las probabilidades de la distribución binomial para $n = 12$ y $p = 0.50$, las cuales se reproducen en la tabla 19.1. La figura 19.1 es la gráfica de esta distribución muestral binomial. En esta tabla se presenta la probabilidad para cada número de signos más, bajo la suposición de que H_0 es verdadera. Ahora se realiza la prueba para determinar si hay diferencia en las preferencias del público por estas marcas de jugos de naranja. Como nivel de significancia se usará 0.05.

En la tabla 19.2 se presentan los datos obtenidos sobre la preferencia. Los dos signos más, indican que dos consumidores prefirieron Citrus Valley. Ahora se pueden usar las distribuciones binomiales para determinar el valor-p de la prueba. Como es una prueba de dos colas, el valor-p se encuentra al duplicar la probabilidad en una cola de la distribución muestral binomial. El nú-

FIGURA 19.1 DISTRIBUCIÓN MUESTRAL BINOMIAL PARA EL NÚMERO DE SIGNOS MÁS CON $n = 12$ Y $p = 0.50$



mero de signos más para Sun Coast Farms (2) se encuentra en la cola inferior de la distribución. De manera que la probabilidad en esa cola es la probabilidad de 2, 1 y 0 signos más. Al sumar estas probabilidades se obtiene $0.0161 + 0.0029 + 0.0002 = 0.0192$. Al duplicar este valor se obtiene el valor-$p = 2(0.0192) = 0.0384$. Como el valor-$p \leq \alpha = 0.05$, se rechaza H_0. Esta prueba de sabores proporciona evidencias de que las preferencias de los consumidores difieren significativamente entre estas dos marcas de jugo de naranja. Se le informará a Sun Coast Farms que los consumidores prefieren Tropical Orange.

La prueba de hipótesis para Sun Coast Farms fue una prueba de dos colas. Como resultado el valor-p se halló al duplicar la probabilidad en una de las colas de la distribución binomial. También se puede hacer una prueba de signo de una cola. Si la prueba es de la cola inferior, el valor-p es la probabilidad de que el número de signos más sea menor o igual al número observado. Si la prueba es de la cola superior, el valor-p es la probabilidad de que el número de signos más sea mayor o igual al número observado.

TABLA 19.2 DATOS DE PREFERENCIAS EN LA PRUEBA DE SUN COAST FARMS

Individuo	Marca preferida	Dato registrado
1	Tropical Orange	−
2	Tropical Orange	−
3	Citrus Valley	+
4	Tropical Orange	−
5	Tropical Orange	−
6	Tropical Orange	−
7	Tropical Orange	−
8	Tropical Orange	−
9	Citrus Valley	+
10	Tropical Orange	−
11	Tropical Orange	−
12	Tropical Orange	−

En la prueba de sabores de Sun Coast Farms, los 12 individuos establecieron su preferencia por una de las dos marcas de jugo de naranja. En otras aplicaciones de la prueba de los signos, puede ocurrir que uno o más de los individuos de la muestra no puedan establecer su preferencia. Si una preferencia no puede ser establecida, se descarta de la muestra esa respuesta y la prueba de los signos se basará en la muestra de menor tamaño. Por último, las probabilidades binomiales que se presentan en la tabla 5 del apéndice B pueden usarse para pruebas de los signos con tamaños de muestras hasta $n = 20$. Para tamaños de muestra mayores, se usa la aproximación normal de las probabilidades binomiales.

Caso de muestras grandes

La prueba de los signos con muestras grandes es equivalente a una proporción poblacional con p = 0.50, como las presentadas en el capítulo 9.

Si la hipótesis nula es H_0: $p = 0.50$ y el tamaño de la muestra es $n > 20$, la distribución muestral del número de signos más se aproxima mediante una distribución normal.

APROXIMACIÓN NORMAL DE LA DISTRIBUCIÓN MUESTRAL DEL NÚMERO DE SIGNOS MÁS CUANDO H_0: $p = 0.50$

$$\text{Media: } \mu = 0.50n \tag{19.1}$$

$$\text{Desviación estándar: } \sigma = \sqrt{0.25n} \tag{19.2}$$

Forma de la distribución: aproximadamente normal siempre que $n > 20$.

A continuación se considera una aplicación de la prueba de los signos en la que se hace un sondeo político. En un sondeo realizado durante una campaña para elecciones presidenciales se pidió a 200 votantes registrados que evaluaran a los candidatos demócrata y republicano con relación a su política exterior. El resultado obtenido fue: 72 de los encuestados evaluaron mejor al candidato demócrata, 103 evaluaron mejor al republicano y 25 no encontraron diferencia entre los candidatos. ¿Con este sondeo puede observarse que exista una diferencia significativa, entre los candidatos, en términos de la opinión pública acerca de su política exterior?

Los casos de empate se eliminan del análisis

Se tiene que $n = 200 - 25 = 175$ fueron las personas que pudieron indicar qué candidato consideraban que tenía una mejor política exterior. Mediante la prueba de los signos y las ecuaciones (19.1) y (19.2) se puede hallar que la distribución muestral del número de signos más tiene las propiedades siguientes.

$$\mu = 0.50n = 0.50(175) = 87.5$$
$$\sigma = \sqrt{0.25n} = \sqrt{0.25(175)} = 6.6$$

Además, como $n = 175$, se puede asumir que la distribución muestral es aproximadamente normal. En la figura 19.2 se muestra esta distribución.

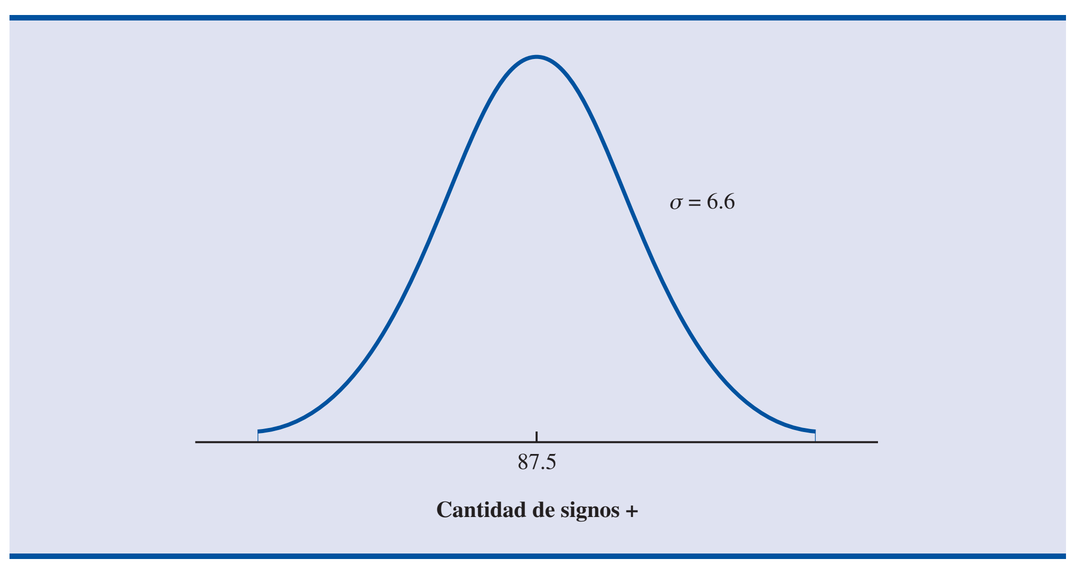
Ahora se procede a realizar la prueba de los signos con un nivel de significancia de 0.05, para obtener las conclusiones. Con base en el número de signos más ($x = 72$) que corresponden al número de personas que evaluaron como mejor la política exterior del candidato demócrata, se obtiene el valor siguiente para el estadístico de prueba

$$z = \frac{x - \mu}{\sigma} = \frac{72 - 87.5}{6.6} = -2.35$$

Si se usa el número de personas que evaluaron mejor al candidato republicano, z = 2.35, se llegará al mismo resultado.

En las tablas de probabilidad normal estándar se encuentra que el área en la cola, a la izquierda de $z = -2.35$ es 0.0094. Como se trata de una prueba de dos colas, el valor-p = 2(0.0094) = 0.0188. Como se obtiene que valor-$p \leq \alpha = 0.05$, se rechaza H_0. Como resultado de este estudio se encuentra que los candidatos difieren en términos de la opinión pública acerca de su política exterior.

FIGURA 19.2 DISTRIBUCIÓN DE PROBABILIDAD PARA EL NÚMERO DE SIGNOS MÁS EN UNA PRUEBA DE LOS SIGNOS EN LA QUE $n = 175$



Prueba de hipótesis acerca de la mediana

En el capítulo 9 se describió el uso de las pruebas de hipótesis para inferencias acerca de la media poblacional. Ahora se muestra cómo realizar una prueba de hipótesis acerca de la mediana poblacional. Recuerde que la mediana divide a la población de manera que 50% de los valores son mayores o iguales que la mediana y 50% de los valores son menores o iguales a la mediana. Cuando se utiliza la prueba de los signos se anota un signo más por cada dato muestral que sea mayor al valor de la mediana hipotética y un signo menos por cada dato muestral que sea menor al valor de la mediana hipotética. Los datos iguales al valor de la mediana hipotética, se descartan. Los cálculos en esta prueba de los signos se hacen igual.

Como ejemplo se realiza la siguiente prueba de hipótesis acerca del precio mediano de las casas nuevas en una determinada región.

$$H_0\text{: Mediana} = \$230\,000$$
$$H_a\text{: Mediana} \neq \$230\,000$$

En una muestra de 62 casas, 34 tuvieron un precio mayor al de la mediana, 26 tuvieron un precio menor al de la mediana y el precio de 2 de ellas fue exactamente $230 000.

Mediante las ecuaciones (19.1) y (19.2) con $n = 60$ casas, cuyos precios son diferentes a $230 000, se obtiene

$$\mu = 0.50\,n = 0.50(60) = 30$$
$$\sigma = \sqrt{0.25\,n} = \sqrt{0.25(60)} = 3.87$$

Como el número de signos más es $x = 34$, el estadístico de prueba es

$$z = \frac{x - \mu}{\sigma} = \frac{34 - 30}{3.87} = 1.03$$

Al aplicar las tablas de la probabilidad normal estándar con $z = 1.03$, se encuentra que el valor-p para dos colas es $2(1 - 0.8485) = 0.303$. Como el valor-$p > 0.05$, no se puede rechazar H_0. De

acuerdo con los datos muestrales, no es posible rechazar la hipótesis nula que establece que el precio mediano de una casa nueva es $230 000.

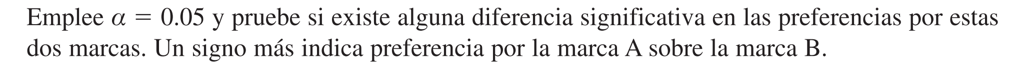
Ejercicios

Métodos



1. En la tabla siguiente se presentan las preferencias de 10 personas respecto a dos marcas de un producto.

Persona	Marca A frente a marca B	Persona	Marca A frente a marca B
1	+	6	+
2	+	7	−
3	+	8	+
4	−	9	−
5	+	10	+



Emplee $\alpha = 0.05$ y pruebe si existe alguna diferencia significativa en las preferencias por estas dos marcas. Un signo más indica preferencia por la marca A sobre la marca B.



2. Realice la prueba de hipótesis siguiente.

$$H_0\text{: Mediana} \leq 150$$
$$H_a\text{: Mediana} > 150$$

En una muestra de tamaño 30 se obtuvieron 22 casos cuyo valor fue mayor que 150, tres cuyo valor fue exactamente 150 y cinco cuyo valor fue menor que 150. Con $\alpha = 0.01$ realice una prueba de hipótesis.

Aplicaciones

3. ¿Las divisiones de acciones son benéficas para los accionistas? La empresa SNL Securities estudió, a lo largo de 18 meses, las divisiones de acciones de la industria de la banca y encontró que las divisiones de las acciones tienden a incrementar el valor de las acciones de un individuo. Admita que en una muestra de 20 recientes divisiones de acciones, 14 hayan llevado a un aumento de su valor, cuatro hayan llevado a una disminución de su valor y dos no hayan ocasionado ningún cambio. Suponga que realiza un estudio para determinar si las divisiones de acciones aún benefician a los poseedores de acciones bancarias.
   a. ¿Cuáles son las hipótesis nula y alternativa?
   b. ¿A qué conclusión se llega con $\alpha = 0.05$?



4. En un sondeo a 1 253 personas se les hizo una serie de preguntas acerca de la economía y del futuro de sus hijos. Una de las preguntas era, "¿Espera que sus hijos tengan una vida mejor a la que usted ha tenido, una vida peor o una vida más o menos igual de buena a la que usted ha tenido?" Las respuestas fueron, 34% mejor, 29% peor, 33% más o menos igual y 4% no supo contestar. Mediante la prueba de los signos y 0.05 como nivel de significancia, determine si el número de adultos que prevén un mejor futuro para sus hijos es mayor al número de adultos que prevén un futuro peor para sus hijos. ¿A qué conclusión llega?
5. La empresa Nielson Media Research identificó a *American Idol* y a *Dancing with the Stars* como los dos programas de televisión de mayor rating en febrero de 2006 (www.nielsenmedia.com, 10 de marzo de 2006). En un estudio local acerca del programa de televisión preferido, de 750 encuestados 330 votaron por *American Idol*, 270 por *Dancing with the Stars* y 150 por otro programa de televisión. Con 0.05 como nivel de significancia pruebe la hipótesis de que *American Idol* y *Dancing with the Stars* tienen el mismo nivel de preferencia. ¿A qué conclusión llega?

6. En el mercado de las computadoras personales la competencia es intensa. En una muestra de 500 compras, se encontró que 202 eran compras de la marca A, 158 de la marca B y 140 de otras marcas. Con un nivel de significancia de 0.05 pruebe la hipótesis de que las marcas A y B tienen la misma participación en el mercado de las computadoras personales. ¿Cuál es la conclusión?
7. El ingreso mediano anual de los suscriptores de la revista *Barron* es $131 000 (barrons.mag.com, 28 de julio de 2000). Suponga que en una muestra de 300 suscriptores a *The Wall Street Journal,* 165 suscriptores posean un ingreso mayor que $131 000 y 135 posean un ingreso menor que $131 000. ¿Puede concluir que hay diferencia entre los ingresos medianos de los dos grupos de suscriptores? Emplee $\alpha = 0.05$ como nivel de significancia, ¿a qué conclusión llega?
8. En una muestra de 150 partidos de básquetbol universitario, el equipo de casa ganó 98 partidos. Realice una prueba para determinar si los datos sustentan la hipótesis de que en el básquetbol universitario el equipo de casa tiene ventaja. ¿A qué conclusión llega con $\alpha = 0.05$?
9. El año pasado, en una determinada ciudad, la mediana del número de empleados de tiempo parcial en un restaurante de comida rápida era 15. Es posible que esta cantidad esté aumentando. En una muestra de nueve restaurantes de comida rápida se encontró que en siete de ellos trabajaban más de 15 empleados de tiempo parcial, en uno había exactamente 15 empleados que trabajaban de tiempo parcial y en otro más había menos de 15 empleados que trabajaban de tiempo parcial. Realice una prueba con $\alpha = 0.05$ para determinar si el número mediano de empleados que trabaja de tiempo parcial ha aumentado.
10. De acuerdo con un estudio nacional, el ingreso anual mediano que los adultos dicen haría realidad sus sueños es $152 000. Suponga que en Ohio, de 225 personas tomadas en una muestra, 122 indican que el ingreso necesario para hacer realidad sus sueños sea menor que $152 000, y 103 informen que esta cantidad sea mayor que $152 000. Pruebe la hipótesis nula de que en Ohio, el ingreso medio anual para que una persona haga realidad sus sueños es $152 000. Use $\alpha = 0.05$. ¿A qué conclusión llega?
11. El ingreso medio anual de los estudiantes con una licenciatura (en Estados Unidos) es $37 700 (*The New York Times Almanac*, 2006). A continuación se presentan los datos muestrales (en miles de dólares) de estudiantes universitarios en la zona de Chicago. Con los datos muestrales pruebe H_0: mediana ≤ 37.7 y H_a: mediana > 37.7 para la población de estudiantes con grado de licenciatura que trabajan en la zona de Chicago. Use $\alpha = 0.05$ como nivel de significancia. ¿Cuál es su conclusión?



Annual

47.8	41.7	31.4	56.9	55.2
47.2	42.6	105.3	38.8	30.0
55.5	127.8	73.7	25.2	68.4
41.2	45.7	37.7	30.4	91.1
21.3	42.4	61.2	23.8	34.1
42.4	25.0	43.2	36.2	76.7
51.9	25.3	39.3	65.0	38.0
32.8	24.4	69.0	25.1	48.7
30.2	60.6	43.4	34.9	37.7
38.5	31.1	91.0	23.6	56.1

19.2 Prueba de los rangos con signo de Wilcoxon

La **prueba de los rangos con signo de Wilcoxon** es la alternativa no paramétrica al método paramétrico de las muestras por pares (o apareadas) presentado en el capítulo 10. En la situación de las muestras por pares, cada unidad experimental genera dos observaciones, una correspondiente a la población 1 y otra correspondiente a la población 2. Las diferencias entre los pares de observaciones permiten apreciar la diferencia entre las dos poblaciones.

En una fábrica se desea determinar cuál de dos métodos de producción difiere en el tiempo que se requiere para realizar una tarea. Se selecciona una muestra de 11 trabajadores y cada trabajador realiza la tarea con uno de estos dos métodos de producción. El método de producción

TABLA 19.3 TIEMPO EN MINUTOS PARA LA REALIZACIÓN DE UNA TAREA DE PRODUCCIÓN

	Método		
Trabajador	**1**	**2**	**Diferencia**
1	10.2	9.5	0.7
2	9.6	9.8	−0.2
3	9.2	8.8	0.4
4	10.6	10.1	0.5
5	9.9	10.3	−0.4
6	10.2	9.3	0.9
7	10.6	10.5	0.1
8	10.0	10.0	0.0
9	11.2	10.6	0.6
10	10.7	10.2	0.5
11	10.6	9.8	0.8

que usa primero cada trabajador es seleccionado de manera aleatoria. De manera que cada uno de los trabajadores de la muestra proporciona un par de observaciones como aparece en la tabla 19.3. Una diferencia positiva entre los tiempos de realización de la tarea indica que el método 1 requiere más tiempo, y una diferencia negativa entre los tiempos indica que el método 2 requiere más tiempo. ¿Los datos obtenidos indican que estos métodos son significativamente diferentes en términos del tiempo que se requiere para realizar la tarea?

En efecto, se tienen dos poblaciones de tiempos requeridos para realizar una tarea, cada población corresponde a cada uno de los métodos; las hipótesis a probar son las siguientes.

H_0: Las poblaciones son idénticas

H_a: Las poblaciones no son idénticas

Si no se puede rechazar H_0, no se contará con evidencia para concluir que los dos métodos difieren en los tiempos requeridos para realizar la tarea. Pero, si H_0 puede ser rechazada, se concluirá que los dos métodos difieren en los tiempos para realizar la tarea.

El primer paso en la prueba de los rangos con signo de Wilcoxon es ordenar los *valores absolutos* de las diferencias entre los dos métodos y asignarles un rango. Toda diferencia que sea igual a cero se descarta y las diferencias restantes se ordenan y se les asigna un rango. A las diferencias que tengan un mismo valor, el rango que se les asigna es el promedio de los números de sus posiciones en el conjunto de datos ordenados. En la última columna de la tabla 19.4 se muestran los rangos asignados a los valores absolutos de las diferencias. Observe que la diferencia cero obtenida por el trabajador 8 se descarta; después, a la diferencia absoluta más pequeña, que es 0.1 se le asigna el rango 1. Se continúa ordenando las diferencias absolutas hasta asignarle a la mayor diferencia absoluta, que es 0.9, el rango 10. El rango que se le asigna a cada una de las diferencias absolutas de los trabajadores 3 y 5, que son iguales, es el promedio de sus posiciones en el conjunto ordenado de las diferencias absolutas 3.5 y el rango para cada una de las diferencias absolutas iguales de los trabajadores 4 y 10 es el promedio de las posiciones que les corresponden en el conjunto ordenado de los datos, 5.5.

Una vez determinados los rangos de las diferencias absolutas, se les antepone el signo de la diferencia original entre los datos. Por ejemplo, a la diferencia 0.1 del trabajador 7, a la que se le ha asignado el rango 1 se le da el valor +1, ya que la diferencia observada entre los dos métodos es positiva. A la diferencia 0.2, que se le asignó el rango 2, se le da el valor −2 ya que la diferencia observada entre los dos métodos es negativa. En la última columna de la tabla 19.4 se encuentra la lista completa de todos los rangos así como la suma de todos ellos.

Ahora se vuelve a la hipótesis original de que las poblaciones de los tiempos necesarios para realizar la tarea, con cada uno de estos dos métodos, son iguales. Si las poblaciones que re-

TABLA 19.4 RANGOS DE LAS DIFERENCIAS ABSOLUTAS ACERCA DEL TIEMPO NECESARIO PARA REALIZAR UNA TAREA DE PRODUCCIÓN

Trabajador	Diferencia	Valor absoluto de la diferencia	Rango	Rango con signo
1	0.7	0.7	8	+ 8
2	−0.2	0.2	2	+ 2
3	0.4	0.4	3.5	+ 3.5
4	0.5	0.5	5.5	+ 5.5
5	−0.4	0.4	3.5	− 3.5
6	0.9	0.9	10	+10
7	0.1	0.1	1	+ 1
8	0.0	0.0	—	—
9	0.6	0.6	7	+ 7
10	0.5	0.5	5.5	+ 5.5
11	0.8	0.8	9	+ 9
			Suma de los rangos con signo	+44.0

presentan los tiempos requeridos para realizar la tarea con cada uno de los métodos fueran idénticas, se esperaría que los rangos positivos y los rangos negativos se compensaran unos con otros y se anularan, de manera que la suma de los valores de los rangos con signo sería aproximadamente cero. Por tanto, en la prueba de los rangos con signo de Wilcoxon, la prueba de significancia consiste en determinar si la suma de los rangos con signo (+44 en este caso) es significativamente distinta de cero.

Sea T la suma de los valores de los rangos con signo en una prueba de los rangos con signo de Wilcoxon. Si las dos poblaciones son idénticas y si el número de pares de datos es 10 o mayor, es posible demostrar que la distribución muestral de T puede ser aproximada mediante una distribución normal.

DISTRIBUCIÓN MUESTRAL DE T PARA POBLACIONES IDÉNTICAS

$$\text{Media: } \mu_T = 0 \tag{19.3}$$

$$\text{Desviación estándar: } \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} \tag{19.4}$$

Forma de la distribución: aproximadamente normal siempre que $n \geq 10$.

En el ejemplo, después de descartar la observación en que la diferencia es cero (la del trabajador 8), se tiene $n = 10$. Por tanto, si emplea la ecuación (19.4), tiene

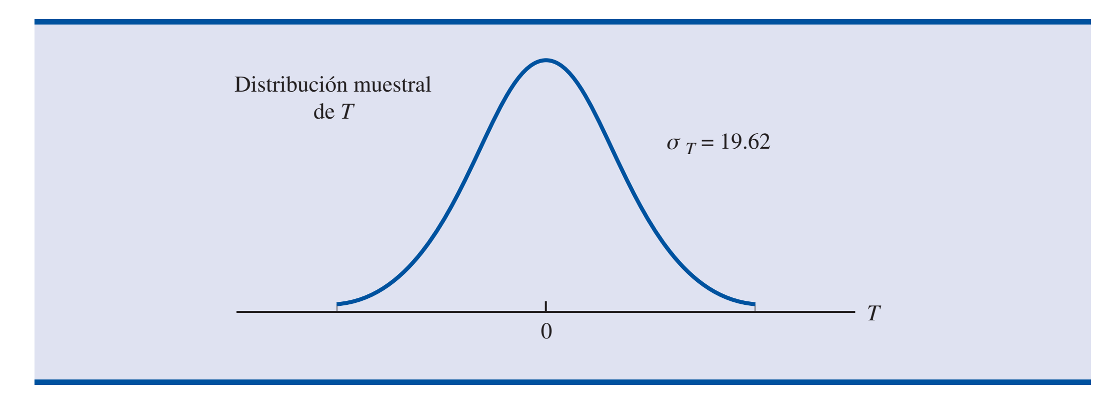
$$\sigma_T = \sqrt{\frac{10(11)(21)}{6}} = 19.62$$

En la figura 19.3 se presenta la distribución muestral de T bajo la suposición de que las dos poblaciones son idénticas. Ahora se procede a realizar la prueba de los rangos con signo de Wilcoxon con 0.05 como nivel de significancia, para llegar a una conclusión. Con la suma de los valores de los signos con rango $T = 44$, se obtiene el valor siguiente para el estadístico de prueba.

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{44 - 0}{19.62} = 2.24$$

A partir de las tablas de probabilidad normal estándar y $z = 2.24$, se halla que para dos colas el valor-$p = 2(1 - 0.9875) = 0.025$. Como el valor-$p \leq \alpha = 0.05$, se rechaza H_0 y se concluye que

FIGURA 19.3 DISTRIBUCIÓN MUESTRAL DE LA T DE WILCOXON DEL TIEMPO PARA LA REALIZACIÓN DE UNA TAREA DE PRODUCCIÓN



las dos poblaciones no son idénticas y que los métodos difieren en el tiempo requerido para realizar la tarea. Como 8 trabajadores obtuvieron tiempos más cortos con el método 2, se concluye que el método 2 es el método de producción que se preferirá.

Ejercicios

Aplicaciones



12. Con objeto de determinar su efecto en el rendimiento de la gasolina en millas por galón en los automóviles de pasajeros, se prueban dos aditivos para gasolina. A continuación aparecen los resultados de esta prueba en 12 automóviles; en cada automóvil se probaron los dos aditivos. Use $\alpha = 0.05$ y la prueba de los rangos con signo de Wilcoxon para determinar si existe una diferencia significativa entre estos dos aditivos.

	Aditivo			Aditivo	
Automóvil	**1**	**2**	**Automóvil**	**1**	**2**
1	20.12	18.05	7	16.16	17.20
2	23.56	21.77	8	18.55	14.98
3	22.03	22.57	9	21.87	20.03
4	19.15	17.06	10	24.23	21.15
5	21.23	21.22	11	23.21	22.78
6	24.77	23.80	12	25.02	23.70



13. Para medir el tiempo que necesitaban para quedarse dormidos, un estudio probó el efecto de un relajante para hombres. Los datos siguientes corresponden a los minutos que requirieron cada uno de los 10 hombres de la muestra para quedarse dormidos. Use como nivel de significancia $\alpha = 0.05$ y determine si el relajante reduce el tiempo que se requiere para quedarse dormido. ¿Cuál es su conclusión?

Sujeto	Sin relajante	Con relajante	Sujeto	Sin relajante	Con relajante
1	15	10	6	7	5
2	12	10	7	8	10
3	22	12	8	10	7
4	8	11	9	14	11
5	10	9	10	9	6

14. En 10 de los principales aeropuertos se muestrearon los precios de la gasolina para automóviles rentados. A continuación se presentan los datos correspondientes a las empresas Avis y Budget (*USA Today,* 4 de abril de 2000).

Aeropuerto	**Avis**	**Budget**
Boston Logan	1.58	1.39
Chicago O'Hare	1.60	1.55
Chicago Midway	1.53	1.55
Denver	1.55	1.51
Fort Lauderdale	1.57	1.58
Los Ángeles	1.80	1.74
Miami	1.62	1.60
Nueva York (JFK)	1.69	1.60
Orange County, CA	1.75	1.59
Washington (Dulles)	1.55	1.54

Use $\alpha = 0.05$ para probar la hipótesis de que no hay diferencia entre las dos poblaciones. ¿Cuál es su conclusión?

15. Dos servicios nocturnos de paquetería fueron probados; se formaron dos muestras idénticas, de manera que a los dos servicios de paquetería se les notificara al mismo tiempo que se requerían sus servicios. A continuación se presentan los tiempos requeridos en cada entrega. ¿Estos datos sugieren que existe diferencia entre los tiempos que requiere cada uno de estos servicios?

	Servicio	
Entrega	**1**	**2**
1	24.5	28.0
2	26.0	25.5
3	28.0	32.0
4	21.0	20.0
5	18.0	19.5
6	36.0	28.0
7	25.0	29.0
8	21.0	22.0
9	24.0	23.5
10	26.0	29.5
11	31.0	30.0

16. El campeonato de los jugadores de la PGA tuvo lugar, del 23 al 26 de marzo de 2006, en el campo de golf TPC Sawgrass en Ponte Vedra Beach, Florida. A continuación se presentan las puntuaciones obtenidas, en la primera y segunda rondas, por 11 golfistas de una muestra. Use $\alpha = 0.05$ y determine si existe una diferencia significativa entre las puntuaciones obtenidas por los golfistas en la primera y en la segunda rondas. ¿Cuál es su conclusión?

Golfista	**Primera ronda**	**Segunda ronda**
Fred Couples	69	73
John Daly	70	73
Ernie Els	72	70
Jim Furyk	65	71
Phil Mickelson	70	73
Rocco Mediate	69	74
Nick Price	72	71
Vijay Singh	68	70
Sergio Garcia	70	68
Mike Weir	71	71
Tiger Woods	72	69

17. Como parte de una investigación de mercado que tenía por objeto evaluar la efectividad de una campaña de publicidad, se seleccionaron 10 ciudades para una prueba de mercado. Las ventas en dólares en cada una de estas ciudades, en la semana anterior a la campaña, se registraron. Después, se realizó la campaña durante dos semanas y se registraron las ventas que hubo en la primera semana, inmediatamente después de la campaña.

Ciudad	Ventas antes de la campaña	Ventas después de la campaña
Kansas City	130	160
Dayton	100	105
Cincinnati	120	140
Columbus	95	90
Cleveland	140	130
Indianapolis	80	82
Louisville	65	55
St. Louis	90	105
Pittsburgh	140	152
Peoria	125	140

Use $\alpha = 0.05$. ¿A qué conclusión llega acerca del valor de la campaña?



19.3 Prueba de Mann-Whitney-Wilcoxon

En esta sección se presenta otro método no paramétrico que se usa para determinar si hay diferencia entre dos poblaciones. Esta prueba, a diferencia de la prueba de los rangos con signo, no se basa en una muestra por pares. Aquí se usan dos muestras independientes, una de cada población. Esta prueba fue creada conjuntamente por Mann, Whitney y Wilcoxon. Algunas veces se le llama *prueba de Mann-Whitney* y otras veces *prueba de la suma de rangos de Wilcoxon.* Las dos versiones de esta prueba, la de Mann-Whitney y la de Wilcoxon son equivalentes. Aquí se le llamará **prueba de Mann-Whitney-Wilcoxon (MWW)**.

La prueba no paramétrica de MWW no requiere que los datos sean de intervalo ni tampoco que las poblaciones estén distribuidas normalmente. El único requisito es que la escala de medición de los datos sea por lo menos ordinal. Después, en lugar de probar las diferencias entre las medias de las dos poblaciones, la prueba de MWW determina si las dos poblaciones son idénticas. Las hipótesis en la prueba de MWW son las siguientes.

H_0: Las dos poblaciones son idénticas

H_a: Las dos poblaciones no son idénticas

Caso de muestras pequeñas

La prueba de MWW para el caso de muestras pequeñas se usa siempre que los tamaños de las muestras de ambas poblaciones sean menores o iguales a 10. El uso de la prueba de MWW para muestras pequeñas se ilustrará mediante un ejemplo sobre la preparación académica de los alumnos de la escuela Johnston. La mayoría de los alumnos de la escuela Johnston provienen de la escuela Garfield o de la escuela Mulberry. La cuestión que desean resolver los directivos de la escuela Johnston es si la población de los estudiantes que provenían de la escuela Garfield es idéntica, en términos de preparación académica, a la población de los estudiantes que provenían de la escuela Mulberry. Las hipótesis son las siguientes.

H_0: Las dos poblaciones son idénticas en términos de preparación académica

H_0: Las dos poblaciones no son idénticas en términos de preparación académica

TABLA 19.5 DATOS DE NIVEL ACADÉMICO

Escuela Garfield		Escuela Mulberry	
Estudiante	**Nivel académico**	**Estudiante**	**Nivel académico**
Fields	8	Hart	70
Clark	52	Phipps	202
Jones	112	Kirkwood	144
Tibbs	21	Abbott	175
		Guest	146

Los directivos de la escuela Johnston toman una muestra aleatoria de cuatro estudiantes provenientes de la escuela Garfield y otra muestra aleatoria de cinco estudiantes provenientes de la escuela Mulberry. De cada uno de los nueve estudiantes tomados para el estudio se registra su actual nivel académico. En la tabla 19.5 se presentan los niveles académicos de estos nueve estudiantes.

El primer paso en la prueba de MWW es *reunir en un solo conjunto* todos los datos y ordenarlos de menor a mayor. Al valor menor (nivel académico 8) se le da el rango 1 y al valor mayor (nivel académico 202) se le da el rango 9. En la tabla 19.6 se presentan los nueve estudiantes con sus rangos y ordenados de acuerdo con ellos.

El paso siguiente es sumar los rangos de cada muestra, por separado. Esto se muestra en la tabla 19.7. En la prueba de MWW se puede usar la suma de cualquiera de las muestras. Aquí se usará la suma de la muestra de los cuatro estudiantes de la escuela Garfield. Esta suma se denota con el símbolo T. De manera que en este ejemplo, $T = 11$.

¿Cuáles son las propiedades de la suma de los rangos en la muestra de Garfield? Puede ocurrir que los cuatro estudiantes en la muestra de Garfield sean los cuatro estudiantes que tengan los primeros rangos en este estudio, si este fuera el caso, $T = 1 + 2 + 3 + 4 = 10$ sería el menor valor que podría tener T, la suma de los rangos. Pero también puede ocurrir que los estudiantes de Garfield fueran los cuatro estudiantes que obtuvieran los últimos rangos, en cuyo caso $T = 6 + 7 + 8 + 9 = 30$ sería el mayor valor que podría tomar T. Por tanto, en la muestra de la escuela Garfield, el valor T estará entre 10 y 30.

Observe que valores de T cercanos a 10 significan que la escuela Garfield tiene los estudiantes significativamente mejores, o con rangos más altos, mientras que valores de T cercanos a 30 significan que la escuela Garfield tiene los estudiantes significativamente peores, o con rangos más bajos. Por tanto, si las dos poblaciones de estudiantes fueran idénticas, en términos de preparación académica, se esperaría que los valores de T fueran aproximadamente iguales al promedio de estos dos valores, o sea $(10 + 30)/2 = 20$.

En la tabla 8 del apéndice B se presentan los valores críticos para el estadístico T en la prueba de MWW para el caso en que los tamaños de ambas muestras son menores o iguales a 10. En esta tabla n_1 corresponde al tamaño de la muestra cuya suma de los rangos se está empleando en la prueba. El valor de T_L se lee directamente en la tabla y el valor de T_U se calcula con la ecuación (19.5).

TABLA 19.6 ESTUDIANTES ORDENADOS POR RANGOS

Estudiante	Nivel académico	Rangos dados a las dos muestras juntas	Estudiante	Nivel académico	Rangos dados a las dos muestras juntas
Fields	8	1	Kirkwood	144	6
Tibbs	21	2	Guest	146	7
Clark	52	3	Abbott	175	8
Hart	70	4	Phipps	202	9
Jones	112	5			

TABLA 19.7 SUMAS DE LOS RANGOS DE LOS ESTUDIANTES PROVENIENTES DE CADA UNA DE LAS ESCUELAS

Estudiantes de la escuela Garfield			Estudiantes de la esucela Mulberry		
Estudiante	**Nivel académico**	**Rango en la muestra**	**Estudiante**	**Nivel académico**	**Rango en la muestra**
Fields	8	1	Hart	70	4
Clark	52	3	Phipps	202	9
Jones	112	5	Kirkwood	144	6
Tibbs	21	2	Abbott	175	8
			Guest	146	7
Suma de los rangos		11			34

Ni los valores de T_L ni los de T_U se encuentran en la zona de rechazo. La hipótesis nula de que las poblaciones son idénticas debe rechazarse sólo si T es estrictamente menor que T_L o estrictamente mayor que T_U.

Por ejemplo, para el nivel de significancia 0.05, en la tabla 8 del apéndice B se encuentra que el valor crítico en la cola inferior para el estadístico de prueba en la prueba de MWW con $n_1 = 4$ (Garfield) y $n_2 = 5$ (Mulberry) es $T_L = 12$. El valor crítico en la cola superior para el estadístico de prueba en la prueba de MWW obtenido con la ecuación (19.5) es

$$T_U = 4(4 + 5 + 1) - 12 = 28$$

En consecuencia, la regla de decisión de esta prueba de MWW indica que la hipótesis nula de que las poblaciones son idénticas puede rechazarse si la suma de los rangos de la primera muestra (Garfield) es menor que 12 o mayor que 28. La regla de rechazo puede expresarse como

$$\text{Rechazar } H_0 \text{ si } T < 12 \text{ o } T > 28$$

Al realizar la prueba con la suma de los rangos de los estudiantes de la escuela Mulberry, se tiene $n_1 = 5$, $n_2 = 4$, $T_L = 17$, $T_U = 33$ y $T = 34$. Como $T > T_U$, se llega también a la misma conclusión, rechazar H_0.

En la tabla 19.7 se ve que $T = 11$. Por tanto, se rechaza la hipótesis H_0 y se concluye que la población de los estudiantes de la escuela Garfield es diferente de la población de los estudiantes de Mulberry en términos de preparación académica. Como los estudiantes de la escuela Garfield obtuvieron las mejores puntuaciones académicas, eso indica que los estudiantes de la escuela Garfield están mejor preparados que los estudiantes de la escuela Mulberry.

Caso de muestras grandes

Cuando los tamaños de las dos muestras son mayores o iguales a 10, para realizar la prueba de MWW se puede usar la aproximación normal de la distribución de T. Para ilustrar este caso de las muestras grandes se empleará un ejemplo del Third National Bank.

El Third National Bank tiene dos sucursales. En la tabla 19.8 se presentan los datos obtenidos de dos muestras aleatorias independientes, una de cada sucursal. ¿Estos datos muestran que son idénticas las poblaciones de saldos en las cuentas de cheques de las dos sucursales?

El primer paso en la prueba de MWW es *reunir en un solo conjunto,* todos los datos y ordenarlos de menor a mayor. En la tabla 19.8, de los 22 datos, se observa que el menor es $750 (sexto elemento de la muestra 2), a este dato se le asigna el rango 1. Al terminar con la asignación de rangos se llega a la lista siguiente.

Saldo ($)	Elemento	Rango asignado
750	6o. de la muestra 2	1
800	5o. de la muestra 2	2
805	7o. de la muestra 1	3

(*continúa*)

Saldo ($)	Elemento	Rango asignado
850	2o. de la muestra 2	4
.	.	.
.	.	.
.	.	.
1195	4o. de la muestra 1	21
1200	3o. de la muestra 1	22

Al ordenar los datos, una vez *reunidos en un solo conjunto*, puede ocurrir que los valores de dos o más datos sean iguales. En este caso, el rango que se le asigna a cada uno es el *promedio* de sus posiciones en el conjunto de los datos ordenados. Por ejemplo, al saldo $945 (octavo elemento de la muestra 1) se le asignará el rango 11. Pero, los siguientes dos datos del conjunto son iguales, su valor es $950 (sexto elemento de la muestra 1 y cuarto elemento de la muestra 2); a cada uno de estos dos datos, que les corresponden las posiciones 12 y 13, el rango que se les asigna es 12.5. Al siguiente dato cuyo valor es $955, continuando con el proceso de asignación de rangos, se le asigna el rango 14. En la tabla 19.9 se presenta el conjunto de datos con los rangos asignados a cada observación.

El paso siguiente en la prueba de MWW es sumar los rangos de cada muestra. Estas sumas se presentan en la tabla 19.9. La prueba se puede basar en la suma de los rangos de cualquiera de las muestras. Aquí se usará la suma de los rangos de la sucursal 1. Así, en este ejemplo, $T = 169.5$.

Dado que los tamaños de las muestras son $n_1 = 12$ y $n_2 = 10$ se puede usar la aproximación normal de la distribución muestral de la suma de los rangos T. La distribución muestral está determinada por las expresiones siguientes.

DISTRIBUCIÓN MUESTRAL DE T PARA POBLACIONES IDÉNTICAS

$$\text{Media: } \mu_T = \tfrac{1}{2} n_1(n_1 + n_2 + 1) \quad (19.6)$$

$$\text{Desviación estándar: } \sigma_T = \sqrt{\tfrac{1}{12} n_1 n_2(n_1 + n_2 + 1)} \quad (19.7)$$

Forma de la distribución: aproximadamente normal siempre que $n_1 \geq 10$ y $n_2 \geq 10$.

TABLA 19.8 SALDOS EN LAS CUENTAS DE DOS SUCURSALES DEL BANCO THIRD NATIONAL BANK

Sucursal 1		Sucursal 2	
Cuenta	**Saldo ($)**	**Cuenta**	**Saldo ($)**
1	1095	1	885
2	955	2	850
3	1200	3	915
4	1195	4	950
5	925	5	800
6	950	6	750
7	805	7	865
8	945	8	1000
9	875	9	1050
10	1055	10	935
11	1025		
12	975		

TABLA 19.9 RANGOS CORRESPONDIENTES A LOS DATOS (REUNIDOS EN UN SOLO CONJUNTO) DE LAS DOS MUESTRAS DEL THIRD NATIONAL BANK

	Sucursal 1			Sucursal 2	
Cuenta	Saldo ($)	Rango	Cuenta	Saldo ($)	Rango
1	1095	20	1	885	7
2	955	14	2	850	4
3	1200	22	3	915	8
4	1195	21	4	950	12.5
5	925	9	5	800	2
6	950	12.5	6	750	1
7	805	3	7	865	5
8	945	11	8	1000	16
9	875	6	9	1050	18
10	1055	19	10	935	10
11	1025	17		Suma de rangos	83.5
12	975	15			
	Suma de rangos	169.5			

Para la sucursal 1 se tiene

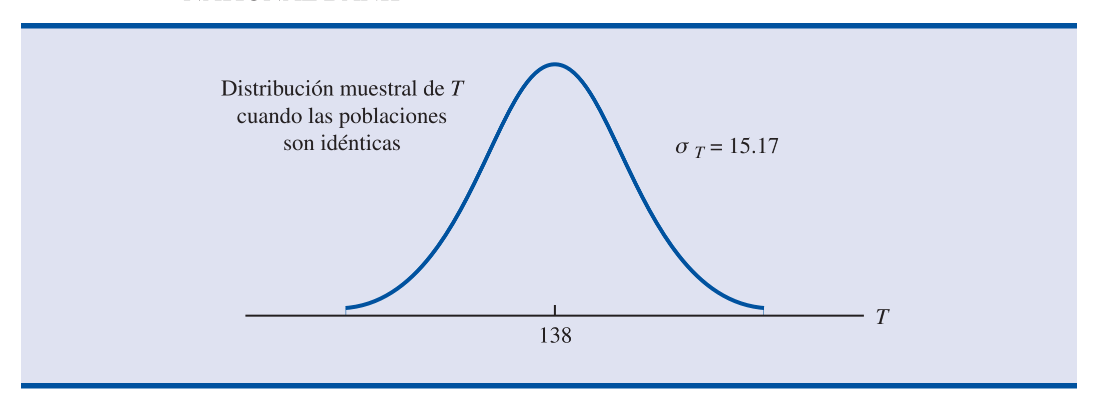
$$\mu_T = \tfrac{1}{2}\,12(12 + 10 + 1) = 138$$
$$\sigma_T = \sqrt{\tfrac{1}{12}\,12(10)(12 + 10 + 1)} = 15.17$$

La figura 19.4 es la distribución muestral de T. Ahora se procede a realizar la prueba de MWW con un nivel de significancia, para llegar a una conclusión, 0.05. Como para la sucursal 1, la suma de los rangos es $T = 169.5$, el valor del estadístico de prueba es el siguiente.

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{169.5 - 138}{15.17} = 2.08$$

En la tabla de la distribución normal estándar, dado que $z = 2.08$, se encuentra que el valor-p para las dos colas es $2(1 - 0.9812) = 0.376$. Como el valor-$p \leq \alpha = 0.05$, se rechaza H_0 y se

FIGURA 19.4 DISTRIBUCIÓN MUESTRAL DE T PARA EL EJEMPLO DE THIRD NATIONAL BANK



concluye que estas dos poblaciones no son idénticas; las poblaciones de los saldos en las cuentas de las dos sucursales no son una misma población.

En resumen, la prueba de la suma de los rangos de Man-Whitney-Wilcoxon para determinar si dos muestras aleatorias independientes pertenecen a poblaciones idénticas consiste en los pasos siguientes.

1. Reunir en un solo conjunto las observaciones muestrales y ordenarlas de menor a mayor al asignarles un rango; a las observaciones muestrales que tengan un mismo valor se les asigna, a cada una, el promedio de los lugares que les corresponden en la lista ordenada de menor a mayor.
2. Calcular T, la suma de los rangos de la primera muestra.
3. En el caso de muestras grandes, para probar si existen diferencias significativas entre las dos poblaciones, el valor obtenido para T se compara con la distribución muestral de T para poblaciones idénticas con las ecuaciones (19.6) y (19.7). Para decidir si se rechaza H_0 se emplea el valor del estadístico de prueba estandarizado z y el valor-p. En el caso de muestras pequeñas, se usa la tabla 9 del apéndice B para hallar los valores críticos para la prueba.

NOTAS Y COMENTARIOS

La prueba no paramétrica vista en esta sección se utiliza para determinar si dos poblaciones son idénticas. Con las pruebas estadísticas paramétricas vistas en el capítulo 10 se prueba la igualdad de dos medias poblacionales. Cuando se rechaza la hipótesis de que las medias sean iguales, se concluye que las poblaciones difieren en sus medias. En la prueba de MWW, cuando se rechaza la hipótesis de que las poblaciones sean idénticas, no se puede decir en qué difieren. Las poblaciones pueden tener diferentes medias, diferentes medianas, diferentes varianzas o diferentes formas. No obstante, si se considera que las poblaciones son iguales en todos los aspectos con excepción de las medias, rechazar H_0 mediante este método no paramétrico implica que las medias son diferentes.

Ejercicios

Aplicaciones



18. Para probar el efecto de dos aditivos sobre el rendimiento de la gasolina, siete automóviles usan el aditivo 1 y nueve automóviles el aditivo 2. En los datos siguientes se presenta el rendimiento en millas por galón obtenido con cada uno de los dos aditivos. Use $\alpha = 0.05$ y la prueba de MWW para determinar si existe una diferencia significativa en el efecto que tienen los dos aditivos sobre el rendimiento.

Aditivo 1	Aditivo 2
17.3	18.7
18.4	17.8
19.1	21.3
16.7	21.0
18.2	22.1
18.6	18.7
17.5	19.8
	20.7
	20.2



19. A continuación se presentan los datos muestrales de los salarios iniciales de contadores públicos y planificadores financieros. Los salarios anuales están dados en miles de dólares.

Contador público	Planificador financiero	Contador público	Planificador financiero
45.2	44.0	50.0	48.6
53.8	44.2	45.9	44.7
51.3	48.1	54.5	48.9
53.2	50.9	52.0	46.8
49.2	46.9	46.9	43.9

a. Use 0.05 como nivel de significancia y pruebe la hipótesis de que no hay diferencia entre los salarios anuales iniciales de los contadores públicos y de los planificadores financieros.
b. Proporcione las medias muestrales de los salarios iniciales en estas dos profesiones.

20. La brecha entre los salarios de hombres y mujeres con la misma preparación disminuye cada vez más, pero aún no se ha cerrado totalmente. A continuación se presentan datos muestrales de siete hombres y siete mujeres con licenciatura. Los datos se dan en miles de dólares.

Hombre	30.6	75.5	45.2	62.2	38.2	49.9	55.3
Mujer	44.5	35.4	27.9	40.5	25.8	47.5	24.8

a. ¿Cuál es el salario mediano de los hombres? ¿Cuál el de las mujeres?
b. Use $\alpha = 0.05$ y realice una prueba de hipótesis para determinar si las dos poblaciones son iguales. Dé su conclusión.

21. Cada año, en diciembre, NRF/BIG Research realiza un estudio sobres el gasto que hacen las personas en las vacaciones de invierno. A continuación se presentan los datos muestrales sobre el gasto en las vacaciones de invierno en 2004 y 2005 (*USA Today*, 20 de diciembre de 2005).

2004	2005
623	752
687	582
748	781
638	805
713	723
645	728
726	674
700	766
794	908
662	737
814	796
674	724

a. Use $\alpha = 0.05$ y realice una prueba para determinar si en 2005 hubo un incremento en comparación con 2004. ¿Cuál es su conclusión?
b. Para cada uno de estos años calcule la media muestral del gasto en vacaciones. Dé el porcentaje en que aumentó o disminuyó el gasto en 2005.

22. *Business Week* publica estadísticas anuales sobre las 1 000 empresas más grandes. El cociente P/E (cociente de rendimiento por acción) de una empresa es el precio actual de las acciones de la empresa dividido entre la ganancia por acción en los últimos 12 meses. En la tabla 19.10 se presenta el cociente P/E de 10 empresas japonesas y 12 empresas estadounidenses de una muestra. ¿Es significativa la diferencia entre los dos países? Use la prueba de MWW y $\alpha = 0.01$ para dar sus conclusiones.

TABLA 19.10 COCIENTE P/E DE ALGUNAS EMPRESAS JAPONESAS Y ESTADOUNIDENSES

Japón		Estados Unidos	
Empresa	**Cociente P/E**	**Empresa**	**Cociente P/E**
Sumitomo Corp.	153	Gannet	19
Kinden	21	Motorola	24
Heiwa	18	Schlumberger	24
NCR Japan	125	Oracle Systems	43
Suzuki Motor	31	Gap	22
Fuji Bank	213	Winn-Dixie	14
Sumitomo Chemical	64	Ingersoll-Rand	21
Seibu Railway	666	American Electric Power	14
Shiseido	33	Hercules	21
Toho Gas	68	Times Mirror	38
		WellPoint Health	15
		Northern States Power	14

23. Los números de delitos por día reportados a la policía durante el verano y el invierno son los siguientes. Use 0.05, como nivel de significancia, para determinar si existe una diferencia significativa entre verano e invierno, en términos del número de crímenes reportados.

Invierno	Verano
18	28
20	18
15	24
16	32
21	18
20	29
12	23
16	38
19	28
20	18

24. Los hornos de microondas de una determinada marca se venden en Dallas y en San Antonio. Los precios se presentan a continuación. Use $\alpha = 0.05$ y pruebe si los precios en Dallas y en San Antonio son los mismos.

Dallas	San Antonio
445	460
489	451
405	435
485	479
439	475
449	445
436	429
420	434
430	410
405	422
	425
	459
	430

25. La National Association of Home Buiders proporciona datos sobre los más frecuentes proyectos de remodelación. Use la prueba de MWW para determinar si se puede concluir que el costo de remodelación de una cocina difiera del costo de remodelación de una recámara. Use 0.05 como nivel de significancia.

Cocina	Recámara
25 200	18 000
17 400	22 900
22 800	26 400
21 900	24 800
19 700	26 900
23 000	17 800
19 700	24 600
16 900	21 000
21 800	
23 600	



Prueba de Kruskal-Wallis

La prueba de MWW, vista en la sección 19.3 se puede usar para probar si dos poblaciones son idénticas. Kruskal y Wallis extendieron esta prueba a tres o más poblaciones. La hipótesis en la **prueba de Kruskal-Wallis** para $k \geq 3$ poblaciones se expresa como sigue.

H_0: Todas las poblaciones son idénticas
H_a: No todas las poblaciones son idénticas

La prueba de Kruskal-Wallis se basa en el análisis de muestras aleatorias independientes de cada una de las k poblaciones.

Esta prueba es una alternativa a la ANOVA presentada en el capítulo 13, en la que se prueba la igualdad de la media de k poblaciones.

En el capítulo 13 se mostró que el análisis de varianza (ANOVA) suele usarse para probar la igualdad de las medias de tres o más poblaciones. En el ANOVA se requieren datos de intervalo o de razón y se requiere suponer que las k poblaciones tienen una distribución normal.

La prueba no paramétrica de Kruskal-Wallis se puede usar tanto con datos ordinales como con datos de intervalo o de razón. Además, en la prueba de Kruskal-Wallis no es necesario suponer que las poblaciones tienen una distribución normal. De manera que siempre que los datos de $k \geq 3$ poblaciones sean ordinales o siempre que la suposición de que las poblaciones tengan una distribución normal sea cuestionable, la prueba de Kruskal-Wallis proporciona un método estadístico alternativo para probar si las poblaciones son idénticas. Esta prueba de Kruskal-Wallis se demostrará con un ejemplo de selección de empleados.

Los empleados que contrata la empresa Williams Manufacturings para su departamento administrativo provienen de tres universidades. Recién el departamento de personal de la empresa ha empezado a revisar el desempeño anual para determinar si hay diferencia en el desempeño de los empleados provenientes de estas tres universidades. Se cuenta con los datos de muestras independientes de clasificación de acuerdo con su desempeño de siete empleados provenientes de la universidad A, seis de la universidad B y siete de la universidad C. En la tabla 19.11 se presentan estos datos; la calificación de acuerdo con su desempeño se da en una escala de 0 a 100.

TABLA 19.11 CALIFICACIONES DE DESEMPEÑO DADAS A 20 EMPLEADOS DE WILLIAMS

Universidad A	Universidad B	Universidad C
25	60	50
70	20	70
60	30	60
85	15	80
95	40	90
90	35	70
80		75

Suponga que se desea probar si las tres poblaciones son idénticas respecto a las calificaciones por su desempeño. Como nivel de significancia se usará 0.05. El estadístico de prueba de Kruskal-Wallis se basa en la suma de los rangos de cada muestra y se calcula como se indica a continuación.

ESTADÍSTICO DE PRUEBA DE KRUSKAL-WALLIS

$$W = \left[\frac{12}{n_T(n_T + 1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right] - 3(n_T + 1) \tag{19.8}$$

donde

k = número de poblaciones
n_i = número de elementos en la muestra i
$n_T = \Sigma n_i$ = número total de los elementos en todas las muestras
R_i = suma de los rangos de la muestra i

Kruskal y Wallis mostraron que bajo la suposición de la hipótesis nula de que las poblaciones son idénticas, la distribución muestral de W puede ser aproximada por una distribución chi-cuadrada con $k - 1$ grados de libertad. Esta aproximación suele ser aceptable siempre que el tamaño de cada una de las muestras sea mayor o igual a cinco. Si el valor del estadístico de prueba es grande, se rechaza la hipótesis nula de que las poblaciones son idénticas. De manera que se usa una prueba de la cola superior.

En la prueba de Kruskal-Wallis únicamente se usa el rango ordinal de los datos.

Para calcular el valor del estadístico W en esta prueba, primero es necesario ordenar los 20 elementos de menor a mayor y asignarles un rango. Al valor menor de estos datos, 15 y que se encuentra en la muestra de la universidad B, se le asigna el rango 1, mientras que al valor mayor, 95 y que se encuentra en la muestra de la universidad A, se le asigna el rango 20. En la tabla 19.12 se presentan los valores de estos datos, sus rangos correspondientes y la suma de los rangos de cada una de las tres muestras. Observe que a los elementos que tienen un mismo valor se les ha asignado un rango promedio;* por ejemplo, los valores 60, 70, 80 y 90 se encuentran repetidos.

Los tamaños de las muestras son

$$n_1 = 7 \qquad n_2 = 6 \qquad n_3 = 7$$

y

$$n_T = \Sigma n_i = 7 + 6 + 7 = 20$$

Mediante la ecuación (19.8) se calcula el estadístico W.

$$W = \frac{12}{20(21)} \left[\frac{(95)^2}{7} + \frac{(27)^2}{6} + \frac{(88)^2}{7} \right] - 3(20 + 1) = 8.92$$

En los procesos con computadora que se presentan en el apéndice F, al final del libro, se muestra el uso de Minitab y Excel para calcular el valor-p.

Ahora se emplea la tabla de la distribución chi-cuadrada (tabla 3 del apéndice B) para determinar el valor-p en esta prueba. Con $k - 1 = 3 - 1 = 2$ grados de libertad, se encuentra que para $\chi^2 = 7.378$, el área en la cola superior de la distribución chi-cuadrada es 0.025 y para $\chi^2 = 9.21$, el área en la cola superior de la distribución chi-cuadrada es 0.01. Como $W = 8.92$ se encuentra entre 7.378 y 9.21, se concluye que el área en la cola superior de la distribución está entre 0.025 y 0.01. Dado que se trata de una prueba de la cola superior, se concluye que el valor-p se encuentra entre 0.025 y 0.01. Con Minitab o con Excel se encuentra que el valor-p = 0.0116. Como el valor-$p \leq \alpha = 0.05$, se rechaza H_0 y se concluye que las poblaciones no son idénticas.

*Si hay muchos valores repetidos, es necesario modificar la ecuación (19.8); la fórmula modificada se encuentra en *Practical Nonparametric Statistics* de W. J. Conover.

TABLA 19.12 RANGOS PARA LOS 20 EMPLEADOS DE WILLIAMS

Universidad A	Rango	Universidad B	Rango	Universidad C	Rango
25	3	60	9	50	7
70	12	20	2	70	12
60	9	30	4	60	9
85	17	15	1	80	15.5
95	20	40	6	90	18.5
90	18.5	35	5	70	12
80	15.5			75	14
Suma de los rangos	95		27		88

El desempeño de los administradores difiere significativamente dependiendo de la universidad de que provienen. Además, dado que las calificaciones al desempeño de los empleados que provienen de la universidad B son las más bajas, sería prudente que la empresa dejara de reclutar empleados de la universidad B o por lo menos los evaluara con más cuidado.

NOTAS Y COMENTARIOS

En el ejemplo usado para ilustrar el procedimiento que se sigue en la prueba de Kruskal-Wallis, lo primero que se hizo fue obtener los datos de nivel de intervalo de las calificaciones del desempeño de los empleados. El procedimiento se hubiera podido realizar también si los datos fueran la clasificación ordinal de los 20 empleados. En ese caso, la prueba de Kruskal-Wallis se hubiera aplicado directamente a los datos originales; el paso de obtención de los rangos a partir de la calificación al desempeño se hubiera omitido.

Ejercicios

Métodos



26. Las calificaciones dadas a tres productos por un panel de 15 consumidores son las siguientes.

Producto		
A	**B**	**C**
50	80	60
62	95	45
75	98	30
48	87	58
65	90	57

Use la prueba de Kruskal-Wallis y $\alpha = 0.05$ para determinar si existe una diferencia significativa entre las calificaciones dadas a los tres productos.

27. Para un examen de admisión se evalúan tres programas de preparación. Las calificaciones obtenidas por las 20 personas de una muestra empleada para probar los programas de preparación son las siguientes. Use la prueba de Kruskal-Wallis para determinar si hay una diferencia significativa entre los tres programas de preparación. Use $\alpha = 0.01$.

Programa		
A	**B**	**C**
540	450	600
400	540	630
490	400	580
530	410	490
490	480	590
610	370	620
	550	570

Aplicaciones



28. Para bajar de peso basta con practicar una de las siguientes actividades tres veces por semana durante cuarenta minutos. En la tabla siguiente se muestra la cantidad de calorías que se quema con 40 minutos de cada una de estas actividades. ¿Estos datos indican que exista diferencia en la cantidad de calorías quemadas con cada una de estas actividades? Dé su conclusión.

Natación	Tenis	Andar en bicicleta
408	415	385
380	485	250
425	450	295
400	420	402
427	530	268

29. La revista *Condé Nast Traveler* realiza cada año un estudio para evaluar los 80 principales barcos cruceros del mundo (*Condé Nast Traveler,* febrero de 2006). A continuación se dan las evaluaciones dadas a los cruceros de una muestra de las líneas Holland America, Princess y Royal Caribbean; la evaluación máxima es 100. Use la prueba de Kruskal-Wallis con $\alpha = 0.05$ para determinar si hay diferencia significativa en las evaluaciones de los barcos de las tres líneas.

Holland America		Princess		Royal Caribbean	
Embarcación	**Evaluación**	**Embarcación**	**Evaluación**	**Embarcación**	**Evaluación**
Amsterdam	84.5	Coral	85.1	Adventure	84.8
Maasdam	81.4	Dawn	79.0	Jewel	81.8
Ooterdam	84.0	Island	83.9	Mariner	84.0
Volendam	78.5	Princess	81.1	Navigator	85.9
Westerdam	80.9	Star	83.7	Serenade	87.4

30. Una empresa grande envía a muchos de sus administrativos de primer nivel a un curso sobre habilidades de supervisión. Este curso se ofrece en cuatro centros educativos y la empresa desea determinar si éstos difieren en la calidad de la capacitación que ofrecen. Para lo cual toma una muestra de 20 de los empleados que han asistido a estos cursos y la muestra se ordena de acuerdo con sus habilidades para la supervisión, dando un rango a cada uno de los componentes de la muestra. Los resultados obtenidos se presentan a continuación.

Curso	Rango de acuerdo con sus habilidades como supervisor				
1	3	14	10	12	13
2	2	7	1	5	11
3	19	16	9	18	17
4	20	4	15	6	8

Observe que el supervisor que obtuvo el mejor rango asistió al curso 2 y el supervisor que obtuvo el peor rango asistió al curso 4. Use $\alpha = 0.05$ y realice una prueba para determinar si hay una diferencia significativa entre la capacitación ofrecida por estos cuatro cursos.

31. Los dulces más vendidos tienen muchas calorías. Los datos siguientes muestran el contenido de calorías en muestras de M&M, Kit Kat y Milky Way II. Pruebe si hay una diferencia significativa en el contenido de calorías de estos tres dulces. Emplee como nivel de significancia 0.05, ¿cuál es su conclusión?

M&Ms	Kit Kat	Milky Way II
230	225	200
210	205	208
240	245	202
250	235	190
230	220	180



Correlación de rangos

El coeficiente de correlación de rangos de Spearman es igual al coeficiente de correlación de Pearson, pero se emplea para datos ordinales.

El coeficiente de correlación es una medida de la relación lineal entre dos variables para las cuales se cuenta con datos de intervalo o de razón. En esta sección se estudia una medida de la relación entre dos variables en el caso de datos ordinales. El **coeficiente de correlación por rangos de Spearman** r_s se usa en estos casos.

COEFICIENTE DE CORRELACIÓN DE SPEARMAN

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)} \qquad \textbf{(19.9)}$$

donde

n = número de elementos o individuos a los que se les va a asignar un rango
x_i = rango del elemento i respecto de una variable
y_i = rango del elemento i respecto de la otra variable
$d_i = x_i - y_i$

A continuación se ilustra el uso del coeficiente de correlación por rangos de Spearman mediante un ejemplo. Una empresa desea determinar si las personas que, en el momento de ser contratadas, generaron expectativas de muy buenos vendedores, en realidad han tenido los mejores registros de ventas. Para esto, el gerente de personal revisa cuidadosamente las entrevistas de trabajo, los antecedentes académicos y las cartas de recomendación de 10 de los vendedores de la empresa. Después de esta revisión, ordena a estas 10 personas de acuerdo con su potencial de éxito, y les da un rango con base en la información disponible al momento de contratarlos. A continuación obtiene una lista del número de unidades vendidas por cada una de estas personas en el transcurso de los primeros dos años y los reordena con un rango de acuerdo con su desempeño real en ventas. En la tabla 19.13 se dan los datos relevantes y los dos rangos. La cuestión es-

TABLA 19.13 POTENCIAL DE VENTAS Y VENTAS REALIZADAS EN DOS AÑOS POR 10 VENDEDORES

Vendedor	Rango de acuerdo con su potencial	Ventas en dos años (unidades)	Rango de acuerdo con sus ventas en dos años
A	2	400	1
B	4	360	3
C	7	300	5
D	1	295	6
E	6	280	7
F	3	350	4
G	10	200	10
H	9	260	8
I	8	220	9
J	5	385	2

tadística es si los rangos, de acuerdo con su potencial de ventas al momento de la contratación, coinciden con los rangos de acuerdo con las ventas realizadas durante los dos primeros años.

Con los datos de la tabla 19.13 se calcula el coeficiente de correlación de rangos de Spearman. En la tabla 19.14 se resumen estos cálculos. Es claro que el coeficiente de correlación por rangos 0.73 es positivo. El coeficiente de correlación por rangos de Spearman varía de -1.0 a $+1.0$ y se interpreta igual que un coeficiente de correlación muestral, en que el valor positivo cercano a 1.0 indica una fuerte relación entre los rangos: si un rango crece el otro crece. Las correlaciones por rangos cercanas a -1.0 indican una fuerte relación pero negativa entre los rangos: cuando un rango crece el otro disminuye. El valor $r_s = 0.73$ indica una correlación positiva entre el desempeño potencial y real. Los individuos con un rango alto de potencial tienden a un alto desempeño.

TABLA 19.14 CÁLCULO DEL COEFICIENTE DE CORRELACIÓN POR RANGOS DE SPEARMAN ENTRE EL POTENCIAL DE VENTAS Y EL DESEMPEÑO EN VENTAS

Vendedor	x_i = Rango de acuerdo con el potencial	y_i = Rango de acuerdo con el desempeño en ventas	$d_i = x_i - y_i$	d_i^2
A	2	1	1	1
B	4	3	1	1
C	7	5	2	4
D	1	6	−5	25
E	6	7	−1	1
F	3	4	−1	1
G	10	10	0	0
H	9	8	1	1
I	8	9	−1	1
J	5	2	3	9
				$\Sigma d_i^2 = 44$

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)} = 1 - \frac{6(44)}{10(100 - 1)} = 0.73$$

Prueba de significancia de la correlación de rangos

Hasta aquí se ha visto cómo usar los resultados muestrales para calcular el coeficiente de correlación por rangos. Como ocurre con muchos otros procedimientos estadísticos, se desea emplear los resultados muestrales para hacer inferencias acerca de la correlación por rangos poblacional ρ_S. Para hacer una inferencia acerca de la correlación por rangos poblacionales, se debe probar la hipótesis siguiente.

$$H_0: p_s = 0$$
$$H_a: p_s \neq 0$$

Bajo la hipótesis nula de que no existe correlación entre los rangos ($\rho_S = 0$), los rangos son independientes y la distribución muestral de r_S es la siguiente.

DISTRIBUCIÓN MUESTRAL DE r_S

$$\text{Media: } \mu_{r_s} = 0 \tag{19.10}$$

$$\text{Desviación estándar: } \sigma_{r_s} = \sqrt{\frac{1}{n-1}} \tag{19.11}$$

Forma de la distribución: aproximadamente normal, siempre que $n \geq 10$.

El coeficiente de correlación por rangos muestrales entre el potencial de ventas y el desempeño en ventas es $r_S = 0.73$. Con este valor se puede probar si hay una correlación por rangos significativa. De acuerdo con la ecuación (19.10) se tiene que $\mu_{r_s} = 0$ y de acuerdo con la ecuación (19.11) se tiene que $\sigma_{r_s} = \sqrt{1/(10-1)} = 0.33$. Si usa como estadístico de prueba la variable aleatoria normal estándar z, tiene

$$z = \frac{r_s - \mu_{r_s}}{\sigma_{r_s}} = \frac{0.73 - 0}{0.33} = 2.20$$

En las tablas de probabilidad normal estándar, se encuentra que para $z = 2.20$, el valor-p = $2(1 - 0.9861) = 0.0278$. Dado que el valor-$p \leq \alpha = 0.05$ se rechaza la hipótesis nula de que la correlación de los rangos sea cero. Por tanto, se puede concluir que hay una correlación de rangos significativa entre el potencial de ventas y el desempeño en ventas.

Ejercicios

Métodos



32. Considere los siguientes conjuntos de rangos dados a los 10 elementos de una muestra.

Elemento	x_i	y_i	Elemento	x_i	y_i
1	10	8	6	2	7
2	6	4	7	8	6
3	7	10	8	5	3
4	3	2	9	1	1
5	4	5	10	9	9

a. Calcule el coeficiente de correlación por rangos de Spearman.
b. Use $\alpha = 0.05$ y pruebe la significancia de la correlación por rangos. Dé su conclusión.

33. Considere los siguientes seis conjuntos de rangos dados a seis objetos.

	Caso uno			Caso dos	
Objeto	Primer rango	Segundo rango	Objeto	Primer rango	Segundo rango
A	1	1	A	1	6
B	2	2	B	2	5
C	3	3	C	3	4
D	4	4	D	4	3
E	5	5	E	5	2
F	6	6	F	6	1

Observe que en el primer caso los rangos son idénticos, mientras que en el segundo los rangos son exactamente opuestos. ¿Cuál es el valor que esperaría para el coeficiente de correlación por rangos de Spearman en cada caso? Explique. Para cada caso calcule el coeficiente de correlación por rangos.

Aplicaciones



34. En la tabla siguiente se presentan los rangos dados para una muestra de 11 estados de acuerdo con el cociente alumnos-profesor (1 = más bajo, 11 más alto) y con los desembolsos por alumno (1 = más alto, 11 más bajo).

	Rango			Rango	
Estado	Cociente alumnos-profesor	Desembolso por alumno	Estado	Cociente alumnos-profesor	Desembolso por alumno
Arizona	10	9	Massachusetts	1	1
Colorado	8	5	Nebraska	2	7
Florida	6	4	North Dakota	7	8
Idaho	11	2	South Dakota	5	10
Iowa	4	6	Washington	9	3
Louisiana	3	11			

Emplee como nivel de significancia $\alpha = 0.05$, ¿parece haber relación entre el desembolso por alumno y el cociente alumnos-profesor?

35. En un estudio realizado por Harris Interactive, Inc. se evaluaron las principales empresas de Internet y se evaluó también su reputación. En la lista siguiente se muestra el ranking de 10 empresas de Internet en relación, por un lado, con su reputación y, por otro, con el porcentaje de entrevistados que dijeron estar dispuestos a comprar acciones de esa empresa.

	Reputación	Probable compra
Microsoft	1	3
Intel	2	4
Dell	3	1
Lucent	4	2
Texas Instruments	5	9
Cisco Systems	6	5
Hewlett-Packard	7	10
IBM	8	6
Motorola	9	7
Yahoo	10	8

a. Calcule la correlación por rangos entre reputación y probable compra.
b. Haga una prueba para determinar si existe una correlación por rangos positiva y significativa. ¿Cuál es el valor-p?
c. Emplee como nivel de significancia 0.05, ¿cuál es su conclusión?

36. A continuación se presenta el ranking de una muestra de golfistas profesionales respecto a "driving distance" y "putting" ¿Cuál es la correlación por rangos entre "driving distance" y "putting"? Como nivel de significancia emplee $\alpha = 0.10$.

Golfista profesional	Driving Distance	Putting
Fred Couples	1	5
David Duval	5	6
Ernie Els	4	10
Nick Faldo	9	2
Tom Lehman	6	7
Justin Leonard	10	3
Davis Love III	2	8
Phil Mickelson	3	9
Greg Norman	7	4
Mark O'Meara	8	1

37. En una determinada universidad, una organización de estudiantes entrevista tanto a estudiantes como a recién egresados para obtener información acerca de la calidad de la enseñanza. Al analizar las respuestas se llega a la siguiente clasificación de los profesores de acuerdo con su habilidad para la enseñanza. ¿Coincide la clasificación dada por los estudiantes con la clasificación dada por los recién egresados? Use $\alpha = 0.10$ y pruebe la significancia de la correlación por rangos.

	Clasificación de acuerdo con	
Profesor	Estudiantes	Recién egresados
1	4	6
2	6	8
3	8	5
4	3	1
5	1	2
6	2	3
7	5	7
8	10	9
9	7	4
10	9	10

Resumen

En este capítulo se presentaron varios métodos estadísticos que se clasifican como métodos no paramétricos. Debido a que los métodos no paramétricos pueden aplicarse a datos nominales y ordinales, así como a datos de intervalo o de razón y a que no se requieren suposiciones acerca de la distribución de la población, estos métodos amplían la clase de problemas que pueden ser sometidos al análisis estadístico.

La prueba de los signos es un método no paramétrico para identificar diferencias entre dos poblaciones, cuando los datos de que se dispone son datos nominales. En el caso de las muestras pequeñas, para determinar los valores críticos de los signos se emplea la distribución de proba-

bilidad binomial; en el caso de las muestras grandes, se emplea una aproximación normal. La prueba de los rangos con signo de Wilcoxon es un método que se emplea para analizar pares de datos, siempre y cuando los datos de cada par sean datos de intervalo o de razón; no es necesario hacer suposiciones acerca de la distribución de la población. Con el método de Wilcoxon se prueba la hipótesis de que las dos poblaciones que se comparan son idénticas.

La prueba de Mann-Whitney-Wilcoxon es un método no paramétrico que se usa para probar la diferencia entre dos poblaciones con base en dos muestras aleatorias independientes; se presentaron tablas para el caso de muestras pequeñas y para el caso de muestras grandes se empleó la aproximación normal. La prueba de Kruskal-Wallis es el análogo no paramétrico a la prueba paramétrica ANOVA para las diferencias entre medias poblacionales.

En la última sección de este capítulo se presentó el coeficiente de correlación por rangos de Spearman, que es una medida de la relación entre dos conjuntos de elementos ordinales o datos ordenados por rangos.

Glosario

Métodos no paramétricos Métodos estadísticos que requieren pocas o ninguna suposición acerca de la distribución de probabilidad de la población y acerca del nivel de medición. Estos métodos suelen usarse cuando se cuenta con datos nominales y ordinales.

Prueba de los signos Prueba estadística no paramétrica para identificar diferencias entre dos poblaciones con base en el análisis de datos nominales.

Prueba de los rangos con signo de Wilcoxon Prueba estadística no paramétrica para identificar diferencias entre dos poblaciones con base en el análisis de dos muestras pareadas.

Prueba de Mann-Whitney-Wilcoxon (MWW) Prueba estadística no paramétrica para identificar diferencias entre dos poblaciones con base en el análisis de dos muestras independientes.

Prueba de Kruskal-Wallis Prueba no paramétrica para identificar diferencias entre tres o más poblaciones.

Coeficiente de correlación por rangos de Spearman Medida de la correlación que se basa en los datos ordenados por rangos de dos variables.

Fórmulas clave

Prueba de los signos (muestras grandes)

$$\text{Media: } \mu = 0.50n \tag{19.1}$$

$$\text{Desviación estándar: } \sigma = \sqrt{0.25n} \tag{19.2}$$

Prueba de los rangos con signo de Wilcoxon

$$\text{Media: } \mu_T = 0 \tag{19.3}$$

$$\text{Desviación estándar: } \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} \tag{19.4}$$

Prueba de Mann-Whitney-Wilcoxon (muestras grandes)

$$\text{Media: } \mu_T = \tfrac{1}{2} n_1(n_1 + n_2 + 1) \tag{19.6}$$

$$\text{Desviación estándar: } \sigma_T = \sqrt{\tfrac{1}{12} n_1 n_2 (n_1 + n_2 + 1)} \tag{19.7}$$

Estadístico de prueba Kruskal-Wallis

$$W = \left[\frac{12}{n_T(n_T + 1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right] - 3(n_T + 1) \tag{19.8}$$

Coeficiente de correlación por rangos de Spearman

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{19.9}$$

Ejercicios complementarios

38. En una encuesta se hizo la siguiente pregunta: ¿usted está a favor o en contra de proporcionar vales libres de impuestos o deducciones de impuestos a los padres que envían a sus hijos a escuelas privadas? De los 2 010 encuestados, 905 estuvieron a favor, 1 045 estuvieron en contra y 60 no tuvieron ninguna opinión al respecto. ¿Estos datos indican que existe una diferencia significativa entre las opiniones respecto al apoyo a los padres que envían a sus hijos a escuelas privadas? Como nivel de significancia use 0.05.

39. El precio mediano en Estados Unidos de una vivienda nueva unifamiliar es $230 000 (The Associated Press, 25 de marzo de 2006). Suponga que los siguientes datos son de las ventas de casas unifamiliares ya existentes en Houston y Boston.

	Más que $230 000	Igual que $230 000	Menos que $230 000
Houston	11	2	32
Boston	27	1	13

a. ¿Es el precio mediano de venta en Houston inferior a la mediana estadounidense? Use una prueba estadística con $\alpha = 0.05$ para respaldar su conclusión.

b. ¿Es el precio mediano de venta en Boston superior a la mediana estadounidense? Use una prueba estadística con $\alpha = 0.05$ para respaldar su conclusión.

40. A 12 amas de casa se les pidió que estimaran el precio de venta de dos modelos de refrigeradores. En la tabla siguiente se muestran las estimaciones que dieron. Use estos datos y 0.05 como nivel de significancia y haga una prueba para determinar si existe alguna diferencia entre los dos modelos en términos de la percepción que tienen las amas de casa sobre sus precios.

Ama de casa	Modelo 1	Modelo 2	Ama de casa	Modelo 1	Modelo 2
1	$650	$900	7	$700	$890
2	760	720	8	690	920
3	740	690	9	900	1000
4	700	850	10	500	690
5	590	920	11	610	700
6	620	800	12	720	700

41. Un estudio busca evaluar la ganancia potencial de peso con cierto alimento aviar. Una muestra de 12 aves se usó durante un periodo de seis semanas. Cada una de las aves se pesó antes y después de un periodo de prueba. Las diferencias entre los pesos antes y después observadas en las 12 aves son: 1.5, 1.2, −0.2, 0, 0.5, 0.7, 0.8, 1.0, 0, 0.6, 0.2, −0.01. Valores negativos indican pérdida de peso durante el periodo de prueba y los ceros indican que no hubo ninguna variación durante el

periodo de prueba. Use 0.05 como nivel de significancia y determine si este nuevo alimento parece ocasionar un aumento de peso en las aves.

42. Los datos siguientes son pesos de un producto en dos líneas de producción. Use $\alpha = 0.05$ y haga una prueba para determinar si existe diferencia entre los pesos de las dos líneas de producción.

Línea de producción 1	**Línea de producción 2**
13.6	13.7
13.8	14.1
14.0	14.2
13.9	14.0
13.4	14.6
13.2	13.5
13.3	14.4
13.6	14.8
12.9	14.5
14.4	14.3
	15.0
	14.9

43. Un cliente desea saber si hay una diferencia significativa entre los tiempos que se requieren para realizar un programa de evaluación por tres métodos diferentes. A continuación se presentan los tiempos (en horas) requeridos por cada uno de los 18 evaluadores para llevar a cabo el programa de evaluación.

Método 1	**Método 2**	**Método 3**
68	62	58
74	73	67
65	75	69
76	68	57
77	72	59
72	70	62

Use $\alpha = 0.05$ y realice una prueba para ver si existe una diferencia significativa entre los tiempos requeridos por los tres métodos.

44. Una muestra de 20 ingenieros que han estado empleados en una empresa durante tres años ha sido ordenada por rangos de acuerdo al potencial administrativo. Algunos de estos ingenieros han asistido a cursos de desarrollo dados por la empresa, otros han asistido a cursos de desarrollo dados por la universidad y los restantes no han asistido a ningún tipo de curso. Emplee $\alpha = 0.025$ y realice una prueba para ver si existe una diferencia significativa entre el potencial administrativo de los tres grupos.

Ningún curso	**Curso de la empresa**	**Curso de la universidad**
16	12	7
9	20	1
10	17	4
15	19	2
11	6	3
13	18	8
	14	5

45. A continuación se presentan las calificaciones dadas en la evaluación a cuatro profesores. Use $\alpha = 0.05$ y el método de Kruskal-Wallis para probar si existe diferencia significativa en sus evaluaciones.

Profesor	**Calificación en la evaluación del curso**								
Black	88	80	79	68	96	69			
Jennings	87	78	82	85	99	99	85	94	
Swanson	88	76	68	82	85	82	84	83	81
Wilson	80	85	56	71	89	87			

46. Los 15 alumnos de una muestra obtuvieron los rangos siguientes en el examen de mitad de semestre y en el examen final de un curso de estadística.

Rango		**Rango**		**Rango**	
Mitad	**Final**	**Mitad**	**Final**	**Mitad**	**Final**
1	4	6	2	11	14
2	7	7	5	12	15
3	1	8	12	13	11
4	3	9	6	14	10
5	8	10	9	15	13

Calcule el coeficiente de correlación por rangos de Spearman y emplee $\alpha = 0.10$, pruebe si hay una correlación significativa.



CAPÍTULO 20

Métodos estadísticos para el control de calidad

CONTENIDO



LA ESTADÍSTICA *en* LA PRÁCTICA

DOW CHEMICAL COMPANY*
FREEPORT, TEXAS

En 1940, la empresa Dow Chemical adquirió un terreno de 800 acres en Texas, en la costa del Golfo, para construir una planta de producción de magnesio. La planta original se ha extendido hasta cubrir más de 5 000 acres y engloba uno de los complejos petroquímicos más grandes del mundo. Entre los productos de Dow Texas Operations se encuentran magnesio, estireno, plásticos, adhesivos, solventes, glicol y cloro. Algunos de los productos se obtienen únicamente para usarlos en otros procesos, pero muchos terminan como ingredientes esenciales de productos farmacéuticos, pastas dentales, alimentos para perros, mangueras, refrigeradores, envases de cartón para leche, bolsas para basura, champús y muebles.

Dow's Texas Operations produce más del 30% del magnesio del mundo; el magnesio es un metal extremadamente ligero que se emplea en productos tan diversos como raquetas de tenis y rines de magnesio. El Departamento del Magnesio fue el primer grupo de Texas Operations que capacitó a su personal técnico y a sus administrativos para usar el control estadístico de calidad. Algunos de los primeros acertados empleos del control estadístico de calidad fueron en los procesos químicos.

En una aplicación en la que intervenía la operación de un secador, se tomaban muestras del producto a intervalos regulares, se calculaba el promedio de cada muestra y se registraba en una carta $\bar{x}$. Estas cartas permitían a los analistas de Dow vigilar tendencias en los productos que pudieran indicar que el proceso no se estaba desarrollando correctamente. En una ocasión los analistas empezaron a observar que las medias muestrales presentaban valores que no correspondían a un proceso que se desarrollara dentro de los límites previstos. Mediante un examen más cuidadoso de las cartas de control y de la operación misma, los analistas encontraron que las variaciones podían deberse a algún problema relacionado con el operador. Después de capacitar nuevamente al operador, las cartas $\bar{x}$ indicaron una mejoría significativa en la calidad del proceso.



El control estadístico de calidad ha permitido a la empresa Dow Chemical mejorar sus métodos de producción y sus productos. © PR Newswire Dow Chemical USA.

En cualquier parte que Dow aplica el control estadístico de calidad, se logra una mejora de la calidad. Se han logrado ahorros documentados de cientos de miles de dólares por año y continuamente se han descubierto nuevas aplicaciones.

En este capítulo se mostrará cómo elaborar una carta $\bar{x}$ como las empleadas por Dow. Estas cartas son parte del control estadístico de calidad conocido como control estadístico de procesos. También se verán métodos de control de calidad que se usan en situaciones en que la decisión de aceptar o rechazar un conjunto de artículos se basa únicamente en una muestra.

* Los autores agradecen a Clifford B. Wilson, administrador de Magnesium Technical, The Dow Chemical Company, por proporcionar este artículo para *La estadística en la práctica.*

La American Society for Quality (ASQ) define la calidad como "la totalidad de rasgos y características de un producto o servicio relacionados con su capacidad para satisfacer determinadas necesidades". Las organizaciones reconocen que en la actual economía global, para ser competitivas tienen que esforzarse por lograr un alto nivel de calidad. De ahí que cada vez se le dé más importancia a los métodos para el monitoreo y el mantenimiento de la calidad.

En la actualidad el enfoque orientado al cliente, fundamental para las organizaciones de alto desempeño, ha transformado el ámbito de las cuestiones de calidad, de la simple eliminación de defectos en la línea de producción a la elaboración de estrategias corporativas de calidad con base amplia. La ampliación del ámbito de la calidad conduce, de manera natural, al concepto de **calidad total (CT)**.

> La calidad total (CT) es un sistema de gestión enfocado a la persona que busca mejorar cada vez más la satisfacción del cliente con costos reales cada vez más bajos. La CT es un sistema (no un

área aparte o un programa de trabajo) y parte integral de una estrategia de alto nivel; la CT funciona a través de todas las funciones y departamentos, involucra a todos los empleados, desde el de más alto nivel hasta el de más bajo y se extiende hacia adelante y hacia atrás incluyendo a la cadena de proveedores y a la cadena de clientes. La CT hace hincapié en el aprendizaje y en la adaptación al cambio continuo como claves para el éxito de una organización.*

Sin importar la manera en que se implemente en las distintas organizaciones, la calidad total se basa en tres principios fundamentales: centrar la atención en el cliente y todos los implicados; participación y trabajo de equipo de toda la organización, y dar especial atención a la mejora continua y al aprendizaje. En la primera sección de este capítulo se presenta una breve introducción acerca de las tres referencias principales en la gestión de calidad: el Malcolm Baldrige Award, los estándares ISO 9000 y la filosofía Seis sigma. En las dos últimas secciones se estudian dos herramientas estadísticas que se usan para monitorear la calidad: el control estadístico de procesos y el muestreo de aceptación.



20.1 Filosofías y marco de referencia

Después de la Segunda Guerra Mundial el doctor W. Edwards Deming fue consultor de la industria japonesa; a él se le atribuye haber convencido a los altos directivos japoneses de usar el método del control estadístico de calidad.

Dos personas que han tenido una gran influencia en las cuestiones sobre la calidad son el doctor W. Edwards Deming y Joseph Juran. Estos dos personajes colaboraron, poco después de la Segunda Guerra Mundial, en la educación de los japoneses en gestión de calidad. Aunque la calidad es un asunto del trabajo de todo mundo, Deming hizo hincapié en que el poner especial atención a la calidad debía ser asunto de los gerentes y directivos; elaboró una lista de 14 puntos que él consideraba que representaban las responsabilidades clave de los gerentes y directivos. Por ejemplo, Deming decía que los gerentes y directivos debían abandonar la inspección en masa; acabar con la práctica de hacer negocios con base únicamente en el precio; que debía buscarse la mejora continua de todos los procesos de producción y servicios; fomentar un ambiente orientado al trabajo en equipo, y que debían eliminarse los objetivos, eslóganes y metas de trabajo basados en cuotas numéricas de trabajo. Lo más importante era que los directivos debían crear un ambiente de trabajo en el que, en todos los niveles, se mantuviera un compromiso continuo con la calidad y la productividad.

Juran propuso una definición simple de calidad: *adecuación al uso.* El método de calidad de Juran se concentraba en tres procesos: planeación de la calidad, control de la calidad y mejoramiento de la calidad. A diferencia de la filosofía de Deming, que requería un cambio cultural importante en la organización, los programas de Juran tenían como fin mejorar la calidad al trabajar con el sistema organizacional existente. No obstante, las dos filosofías se parecen en que se centran en la necesidad de que los directivos se involucren y hagan hincapié en la necesidad de un mejoramiento continuo, en la importancia de la capacitación y en el uso de técnicas de control de calidad.

Hubo otras muchas personas que tuvieron un papel importante en el movimiento de la calidad. Entre ellos se encuentran Philip B. Crosby, A. V. Feigenbaum, Karou Ishikawa y Genichi Taguchi. En textos especializados, dedicados únicamente al tema de la calidad, se encuentran detalles sobre las contribuciones de cada uno de ellos. Las contribuciones de todos los involucrados en el movimiento de calidad ayudaron a definir el conjunto de las mejores prácticas y llevaron a la creación de numerosos programas de premios y de certificación. Los dos programas más importantes son el Malcolm Baldrige National Quality Award, de Estados Unidos, y el proceso internacional de certificación ISO 9000. En los últimos años ha aumentado también el uso de Seis Sigma, una metodología para el mejoramiento del desempeño organizacional que se basa en la recolección de datos y en el análisis estadístico.

Malcolm Baldrige National Quality Award

El Malcolm Baldrige National Quality Award es entregado por el presidente de Estados Unidos a las organizaciones que apliquen y que se les considere destacadas en siete áreas: liderazgo; planeación estratégica; enfoque al cliente y al mercado; gestión de medición, análisis y conocimiento; especial atención al recurso humano; gestión de procesos y resultados económicos. El congreso

*J. R. Evans y W. M. Lindsay, *The Management and Control of Quality*, 6a. ed. (Cincinnati, OH: South-Western, 2005), pp. 18-19.

El National Institute of Standards and Technology (NIST) dependiente del Departamento de Comercio de Estados Unidos es el que se ocupa del Baldrige National Quality Program. Más información se puede obtener en www.quality.nist.gov.

estableció el programa de este premio en 1987 para reconocer a las organizaciones estadounidenses por sus logros en calidad y desempeño, y para llamar la atención acerca de la importancia de la calidad como ventaja competitiva. Malcolm Baldrige trabajó como secretario de comercio de Estados Unidos, desde 1981 hasta su muerte en 1987; el premio lleva su nombre en su honor.

Desde la ceremonia de entrega de los primeros premios en 1988, el Baldrige National Quality Program ha crecido en estatura y en impacto. Desde 1988 se han distribuido aproximadamente 2 millones de copias de los criterios, a lo que se suman las reproducciones en gran escala hechas por las organizaciones y el acceso electrónico. Por octavo año consecutivo, un índice accionario hipotético, formado por empresas estadounidenses que cotizan en la bolsa y que han recibido el Baldrige Award, supera el Standar & Poor's 500. En el 2003, el "Índice Baldrige" superó al S&P 500 por 4.4 a 1. En la ceremonia de premiación de 2003, Bob Barnett, vicepresidente ejecutivo de Motorola, Inc., dijo, "Ingresamos a este programa de calidad, no con la idea de ganar, sino con el objetivo de obtener la evaluación de los examinadores de Baldrige. Esta evaluación fue cabal, profesional y *clara*... haciendo de ella la consultoría más rentable y con mayor valor agregado que se puede obtener, actualmente, en todo el mundo."

ISO 9000

ISO 9000 es una serie de cinco estándares internacionales publicados en 1987 por la Organización Internacional para la Estandarización (ISO), Génova, Suiza. Las empresas pueden usar estos estándares como ayuda, para determinar lo que se necesita para mantener un sistema de conformidad de calidad eficiente, para garantizar que los instrumentos de medición y de prueba sean calibrados regularmente y, para mantener, y adecuar, un sistema apropiado de documentación. El registro ISO 9000 determina si una empresa cumple con su propio sistema de calidad. En general, el registro ISO 9000 cubre menos del 10% de los criterios del Baldrige Award.

Seis Sigma

Al final de los años ochenta, Motorola advirtió la necesidad de mejorar la calidad de sus productos y servicios; su objetivo fue lograr un nivel de calidad tan bueno que en cada millón de operaciones no se presentaran más de 3.4 errores. A este nivel de calidad se le conoce como el nivel de calidad seis sigma y a la metodología creada para lograr este objetivo de calidad se le conoce como **Seis Sigma**.

Una organización puede emprender dos tipos de proyectos Seis Sigma:

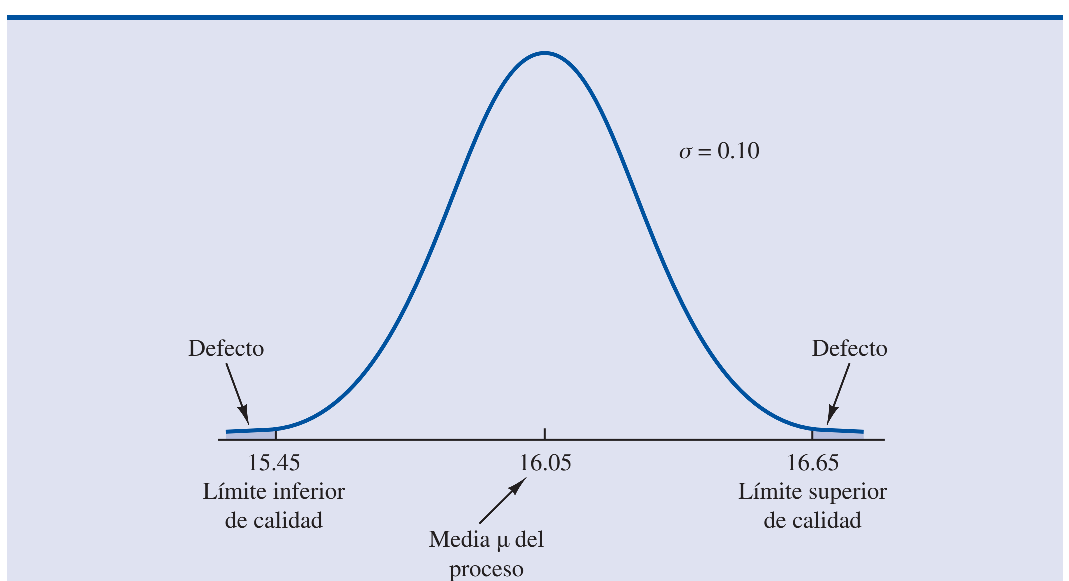
- DMAIC (Definir, Evaluar, Analizar, Mejorar y Controlar) como ayuda para rediseñar procesos ya existentes.
- DFSS (Diseño para Seis sigma) para diseñar nuevos productos, procesos o servicios.

Para ayudar en el rediseño de procesos ya existentes y en el diseño de nuevos procesos, Seis Sigma hace hincapié en el análisis estadístico y en la evaluación cuidadosa. Actualmente, Seis Sigma es una herramienta importante para ayudar a las organizaciones a alcanzar niveles Baldrige de desempeño en negocios y muchos de los examinadores del Baldrige consideran a Seis Sigma como un método ideal para poner en marcha programas de mejoramiento de Baldrige.

Límites de Seis Sigma y defectos por millón de operaciones En la terminología de Seis Sigma, un *defecto* es un error que llega al cliente. El proceso Seis Sigma define desempeño de calidad en términos de defectos por millones de operaciones (dpmo). Como ya se indicó, Seis sigma representa un nivel de calidad de por lo menos 3.4 dpmo. Para ilustrar cómo se evalúa este nivel de calidad se tomará como ejemplo un caso de la empresa empacadora KJW.

KJW tiene una línea de producción en la que llena paquetes de cereal. En este proceso, la media es $\mu = 16.05$ onzas y la desviación estándar es $\sigma = 0.10$ onzas. Suponga que los pesos de llenado siguen una distribución normal. En la figura 20.1 se muestra la distribución de los pesos de llenado. Suponga que los directivos consideran como límites de calidad aceptables para este proceso de 15.45 a 16.65 onzas. Por tanto, todo paquete de cereal que contenga menos que 15.45

FIGURA 20.1 DISTRIBUCIÓN NORMAL DE LOS PESOS DE LLENADO DE LOS PAQUETES DE CEREALES, LA MEDIA ES $\mu = 16.05$

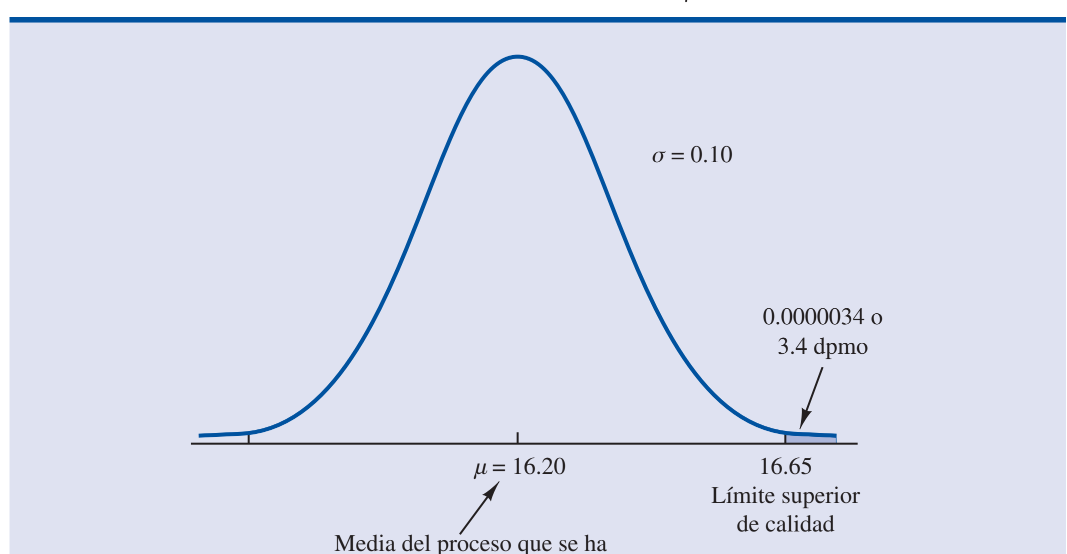


o más que 16.65 onzas será considerado como defecto. Mediante Excel o Minitab se puede mostrar que 99.9999998% de los paquetes llenados pesará entre 16.05 – 6(0.10) = 15.45 onzas y 16.05 + 6(0.10) = 16.65 onzas. En otras palabras, sólo 0.0000002% de los paquetes llenados contendrán menos de 15.45 onzas o más de 16.65 onzas. Por tanto, la posibilidad de que en este proceso de llenado, un paquete de cereal sea defectuoso parece ser extremadamente pequeña, ya que en promedio sólo 2 paquetes de cada 10 millones serán defectuosos.

Motorola se cercioró, en sus primeros trabajos con Seis Sigma, de que la media de un proceso se puede desplazar en promedio hasta 1.5 desviaciones estándar. Por ejemplo, suponga que la media en el proceso de KJW aumente en 1.5 desviaciones estándar, es decir 1.5(0.10) = 0.15 onzas. Con este desplazamiento, la distribución normal de los pesos de llenado tendrá como centro $\mu = 16.05 + 0.15 = 16.20$ onzas. Cuando la media del proceso es $\mu = 16.05$ onzas, la probabilidad de obtener un paquete de cereal que pese más de 16.65 onzas es extremadamente pequeña. Pero, ¿cuál es esta probabilidad si la media se desplaza a 16.20 onzas? En la figura 20.2 se muestra que en este caso, el límite superior de calidad, 16.65 se encuentra a 4.5 desviaciones estándar a la derecha de la nueva media 16.20 onzas. En Excel o Minitab se ve que con esta media, la probabilidad de que un paquete pese más de 16.65 onzas es 0.0000034. Por tanto, si la media del proceso se desplaza hacia arriba 1.5 desviaciones estándar, aproximadamente 1 000 000(0.0000034) = 3.4 paquetes de cereal sobrepasarán el límite superior de 16.65 onzas. En la terminología de Seis Sigma, se dice que el nivel de calidad del proceso es de 3.4 defectos por millón de operaciones. Si los directivos de KJW consideran que 15.45 a 16.65 onzas son límites de calidad aceptables, este proceso de llenado de KJW será considerado un proceso Seis Sigma. Por tanto, si la media del proceso permanece a no más de 1.5 desviaciones estándar de la media deseada $\mu = 16.05$ onzas, se pueden esperar como máximo 3.4 paquetes defectuosos por millón de paquetes llenados.

Las organizaciones que desean alcanzar y mantener un nivel de calidad Seis Sigma deben poner especial cuidado en emplear métodos para la vigilancia y conservación de la calidad. El *aseguramiento de la calidad* se refiere a todo el conjunto de políticas, procedimientos y lineamientos establecidos por la organización para alcanzar y mantener la calidad. El aseguramiento de la calidad consiste en dos funciones principales: ingeniería de calidad y control de calidad. El objeto de la *ingeniería de calidad* es incluir la calidad en el diseño de los productos y procesos e identificar los problemas de calidad antes de la producción. El **control de calidad** consiste en una se-

FIGURA 20.2 DISTRIBUCIÓN NORMAL DE LOS PESOS DE LLENADO DE PAQUETES DE CEREAL, LA MEDIA DEL PROCESO ES $\mu = 16.20$



rie de inspecciones y mediciones usadas para determinar si se están satisfaciendo los estándares de calidad. Si no es el caso, se pueden tomar medidas correctivas o preventivas para alcanzar y mantener la conformidad. En las dos secciones siguientes se presentan dos métodos estadísticos que se usan en el control de calidad. En el primero, el *control estadístico de procesos*, se emplean representaciones graficas conocidas como cartas de control para monitorear un proceso; el objetivo es determinar si puede continuar el proceso o si se deben tomar medidas correctivas para lograr el nivel de calidad deseado. El segundo método, el *muestreo de aceptación,* se emplea cuando la decisión de aceptar o rechazar un conjunto de artículos tiene que basarse en la calidad encontrada en una muestra.



20.2 Control estadístico de procesos

En esta sección se estudiarán procesos de control de calidad que se emplean en los procesos de producción en que los bienes se producen de manera continua. Con base en una muestra y en la inspección del producto del proceso se decide si se puede continuar con el proceso de producción o si es necesario ajustarlo para que los artículos o los bienes producidos estén dentro de los estándares de calidad aceptables.

Uno de los conceptos más importantes en el movimiento de administración de la calidad total es la mejora continua. El principal objetivo de una carta de control es mejorar la calidad.

Aunque se cuente con estándares altos en las operaciones de fabricación y de producción, invariablemente habrá herramientas y maquinaria que se desgastan, vibraciones que hagan que se desajuste la maquinaria, materiales empleados que puedan presentar anomalías y operadores que cometan errores. Todos estos factores suelen dar como resultado un producto de mala calidad. Por fortuna, existen procedimientos para vigilar la calidad de los productos, mediante los cuales se puede detectar oportunamente una mala calidad y ajustar o corregir el proceso de fabricación.

Si las variaciones en la calidad del producto se deben a **causas asignables** como desgaste de las herramientas, ajuste incorrecto de la maquinaria, materia prima de mala calidad o errores del operador, es necesario ajustar o corregir el proceso lo antes posible. Por otro lado, si las variaciones en la calidad del producto se deben a lo que se conoce como **causas comunes** —es decir, variaciones que se presentan de manera aleatoria, como variaciones en la temperatura, la humedad, etc., causas que no puede controlar el fabricante—, no es necesario ajustar el proceso. El objetivo

TABLA 20.1 RESULTADOS DEL CONTROL ESTADÍSTICO DE PROCESOS

		Situación del proceso de producción	
		H_0 es verdadera El proceso está bajo control	**H_0 es falsa El proceso está fuera de control**
Decisión	**Que continúe el proceso**	Decisión correcta	Error tipo II (dejar continuar un proceso que está fuera de control)
	Ajustar el proceso	Error tipo I (ajustar un proceso que se encuentra bajo control)	Decisión correcta

principal del control estadístico de procesos es determinar si las variaciones en el producto se deben a causas asignables o a causas comunes.

Cuando se detectan causas asignables se dice que el proceso está *fuera de control.* En tales casos se toman medidas correctivas para hacer que el proceso regrese a los niveles de calidad aceptables. Si las variaciones que se observan en el producto de un proceso de fabricación se deben únicamente a causas comunes, se concluye que el producto se encuentra *bajo control estadístico* o simplemente *bajo control*; en esos casos no es necesario hacer modificación o ajuste alguno.

Los procedimientos de control de procesos están estrechamente relacionados con los procedimientos de prueba de hipótesis, ya antes vistos en este libro. Las cartas de control permiten realizar sobre la marcha pruebas de la hipótesis de que el proceso está bajo control.

Los métodos estadísticos para el control de procesos se basan en la metodología de las pruebas de hipótesis, presentada en al capítulo 9. La hipótesis nula H_0 se formula considerando que el proceso de producción está bajo control. La hipótesis alternativa H_a se formula considerando que el proceso de producción está fuera de control. En la tabla 20.1 se muestra cómo se toman las decisiones correctas de dejar que continúe un proceso que está bajo control o de ajustar un proceso que está fuera de control. Como ocurre con las demás pruebas de hipótesis, aquí también es posible cometer un error tipo I (ajustar un proceso que está bajo control) o un error tipo II (permitir que continúe un proceso que está fuera de control).

Cartas de control

Las **cartas de control** constituyen la base para decidir si las variaciones en el producto se deben a causas comunes (en control) o causas asignables (fuera de control). Siempre que se detecte que un proceso está fuera de control es necesario realizar ajustes o tomar medidas correctivas que hagan que el proceso regrese a la situación bajo control.

A las cartas de control que se basan en datos que se pueden medir en una escala continua se les llama cartas de control de variables. Las cartas $\bar{x}$ son cartas de control de variables.

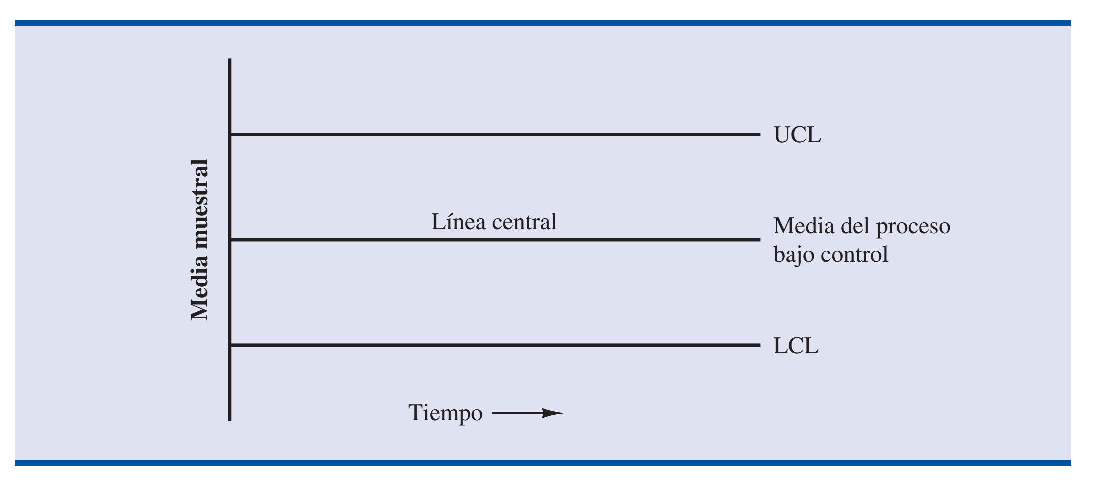
Las cartas de control se clasifican de acuerdo con el tipo de datos que contienen. Se usa una **carta $\bar{x}$** cuando la calidad del producto de un proceso se mide en términos de una variable, como longitud, peso, temperatura, etc. En tal caso, la decisión de dejar continuar el proceso de producción o de ajustarlo se basa en el valor de la media hallada en una muestra del producto. Para introducir algunos de los conceptos que son comunes a todas las cartas de control, se considerarán algunos de los rasgos característicos de una carta $\bar{x}$.

En la figura 20.3 se presenta la estructura general de una carta $\bar{x}$. La línea central que se observa en esta carta corresponde a la media del proceso, cuando el proceso está *bajo control*. La línea vertical identifica la escala de medición para la variable de interés. Cada vez que se toma una muestra del proceso de producción, se obtiene el valor de su media muestral $\bar{x}$ y se grafica el punto correspondiente al valor de $\bar{x}$ en la carta de control.

Las dos líneas rotuladas como UCL y LCL sirven para determinar si el proceso está bajo control o fuera de control. A estas líneas se les llama *límite de control superior* y *límite de control inferior*, respectivamente. Estos límites se eligen de manera que cuando el proceso esté bajo control exista una gran probabilidad de que los valores de $\bar{x}$ estén entre estos dos límites. Si hay valores que estén fuera de los límites de control, éstos serán evidencias estadísticas claras de que el proceso se encuentra fuera de control y que es necesario tomar medidas correctivas.

A medida que pasa el tiempo, se van graficando más y más puntos en la carta de control. El orden en el que se van agregando estos puntos es de izquierda a derecha, a medida que se van toman-

FIGURA 20.3 ESTRUCTURA DE UNA CARTA $\bar{x}$



do las muestras del proceso. En esencia, cada vez que se grafica un nuevo punto en una carta de control, se está realizando una prueba de hipótesis para determinar si el proceso está bajo control.

Además de las cartas $\bar{x}$ se pueden usar otras cartas, como cartas para monitorear el rango de las mediciones en la muestra (**cartas *R***) o para monitorear la proporción de defectos en la muestra (**cartas *p***) o para monitorear la cantidad de defectos en la muestra (**cartas *np***). En todos estos casos, las cartas de control tienen una línea inferior de control (LCL, por sus siglas en inglés), una línea central y una línea superior de control (UCL por sus siglas en inglés) como la carta $\bar{x}$ de la figura 20.3. La principal diferencia entre estas cartas es la que se mide en el eje vertical; por ejemplo, en una carta p la escala de medición, en lugar de denotar la media muestral, denota la proporción de artículos defectuosos existentes en una muestra. A continuación se ilustrará la construcción y el uso de las cartas $\bar{x}$, de las cargas R, de las cartas p y de las cartas np.

Cartas $\bar{x}$: media y desviaciones estándar del proceso conocidas

Para ilustrar la construcción de una carta $\bar{x}$ se empleará el ejemplo de la empacadora KJW. Recuerde que la empresa KJW, cuenta con una de sus líneas de producción en la que llena paquetes de cereal. Cuando el proceso se desarrolla correctamente —y por tanto se encuentra bajo control— el peso medio de llenado es $\mu = 16.05$ onzas y la desviación estándar del proceso es $\sigma = 0.10$ onzas. Se supone, también, que los pesos de llenado siguen una distribución normal. En la figura 20.4 se muestra esta distribución.

Para determinar la variación que puede esperarse en los valores de $\bar{x}$ cuando el proceso está bajo control se usa la distribución muestral de $\bar{x}$, vista en el capítulo 7. Se recuerdan brevemente, las propiedades de la distribución muestral de $\bar{x}$. recuerde que el valor esperado o la media de los valores de $\bar{x}$ es igual a μ, el peso medio de llenado si la línea de producción está bajo control. Si las muestras son de tamaño n, la ecuación para obtener la desviación estándar de $\bar{x}$, que se conoce como error estándar de la media, es

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{(20.1)}$$

Además, como los pesos de llenado están distribuidos normalmente, para cualquier tamaño de muestra, la distribución muestral de $\bar{x}$ es una distribución normal. De manera que la distribución muestral de $\bar{x}$ es una distribución normal con media μ y desviación estándar $\sigma_{\bar{x}}$. En la figura 20.5 se presenta esta distribución.

La distribución muestral de $\bar{x}$ se usa para determinar cuáles son valores razonables de $\bar{x}$ cuando el proceso se halla bajo control. En el control de calidad se suele considerar como razonable todo valor de $\bar{x}$ que no se aleje de la media, hacia arriba o hacia abajo, más de 3 desviaciones estándar o errores estándar. Recuerde que al estudiar la distribución de probabilidad normal se vio

FIGURA 20.4 DISTRIBUCIÓN NORMAL DE LOS PESOS DE LLENADO DE LOS PAQUETES DE CEREAL

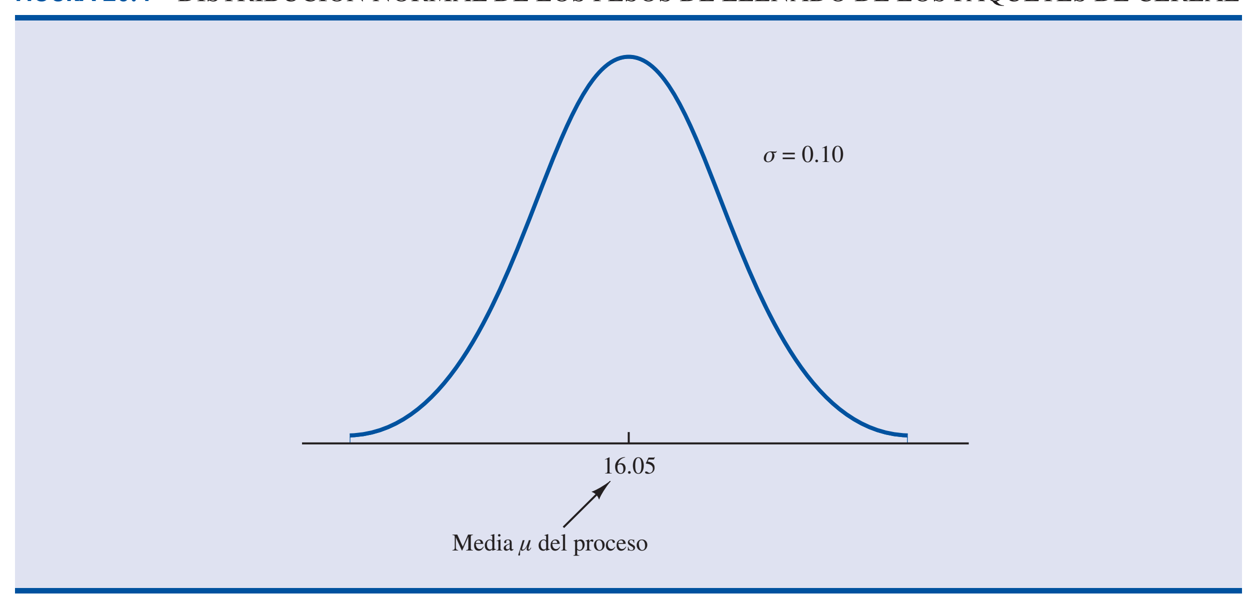
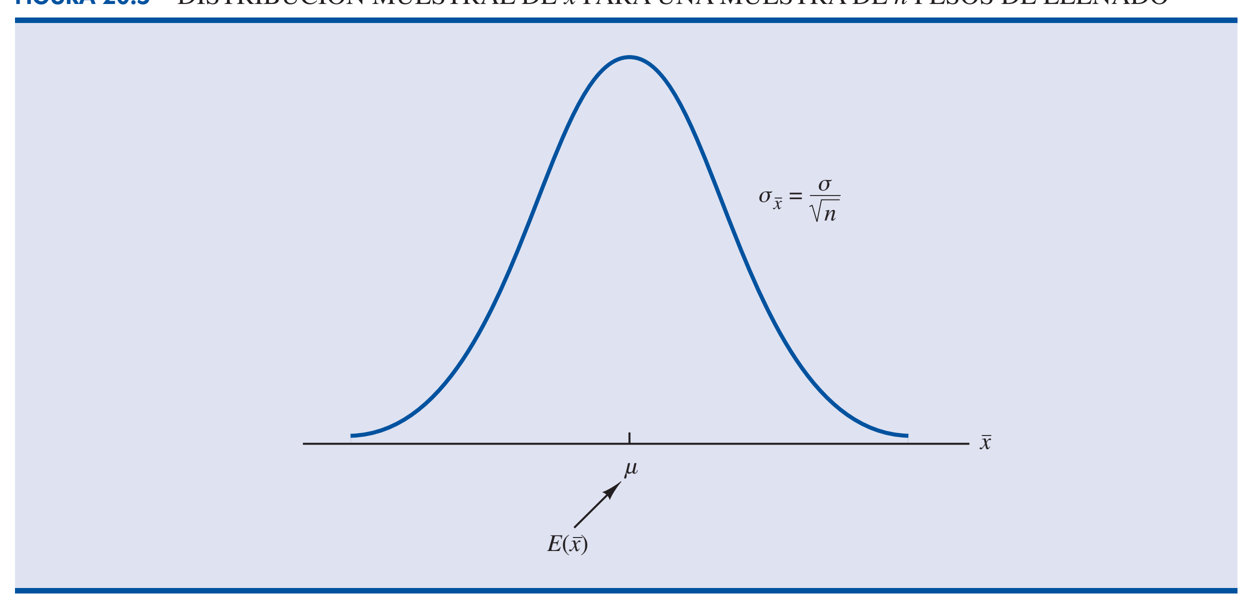


FIGURA 20.5 DISTRIBUCIÓN MUESTRAL DE $\bar{x}$ PARA UNA MUESTRA DE n PESOS DE LLENADO

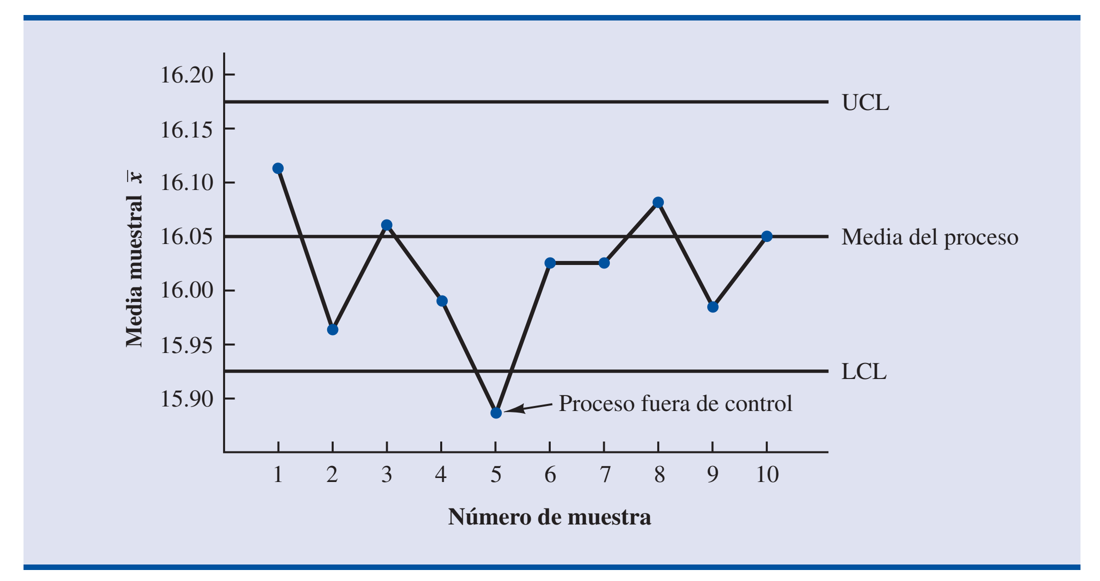


que aproximadamente 99.7% de los valores de una variable aleatoria que tengan una distribución normal se encuentran a no más de ±3 desviaciones estándar de su media. Por tanto, si un valor de $\bar{x}$ se encuentra dentro del intervalo de $\mu - 3\sigma_{\bar{x}}$ a $\mu + 3\sigma_{\bar{x}}$ se admitirá que el proceso está bajo control. En resumen, los límites de control en una carta $\bar{x}$ son los siguientes:

LÍMITES DE CONTROL EN UNA CARTA $\bar{x}$: MEDIA Y DESVIACIÓN ESTÁNDAR DEL PROCESO CONOCIDAS

$$\text{UCL} = \mu + 3\sigma_{\bar{x}} \quad \textbf{(20.2)}$$
$$\text{LCL} = \mu - 3\sigma_{\bar{x}} \quad \textbf{(20.3)}$$

FIGURA 20.6 CARTA $\bar{x}$ DEL PROCESO DE LLENADO DE LOS PAQUETES DE CEREAL



De regreso al ejemplo de la empresa KJW, la distribución de los pesos de llenado se muestra en la figura 20.4 y la distribución muestral de $\bar{x}$ se muestra en la figura 20.5. Suponga que periódicamente se toman seis paquetes del proceso de llenado y se calcula su media muestral para determinar si el proceso está bajo control o fuera de control. Mediante la ecuación (20.1), se encuentra que el error estándar de la media es $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.10/\sqrt{6} = 0.04$. Por tanto, como la media del proceso es 16.05, los límites de control son UCL = 16.05 + 3(0.04) = 16.17 y LCL = 16.05 − 3(0.04) = 15.93. En la figura 20.6 se muestra la carta de control con los resultados de 10 muestras tomadas a lo largo de un periodo de 10 horas. Para facilitar la lectura, los números correspondientes a las muestras se han colocado en la parte inferior de la carta.

Observe que la media de la quinta muestra indica que el proceso está fuera de control. La quinta media muestral se encuentra abajo del LCL (límite de control inferior), lo cual indica que existen causas asignables que ocasionan variación de la calidad del producto y que el llenado se está realizando con una cantidad menor a la estipulada. En este momento, se toman medidas correctivas para hacer que el proceso vuelva a estar bajo control. El hecho de que los demás puntos de la carta $\bar{x}$ se encuentren dentro de los límites de control inferior y superior, indican que la acción correctiva fue adecuada.

Cartas $\bar{x}$: media y desviaciones estándar del proceso desconocidas

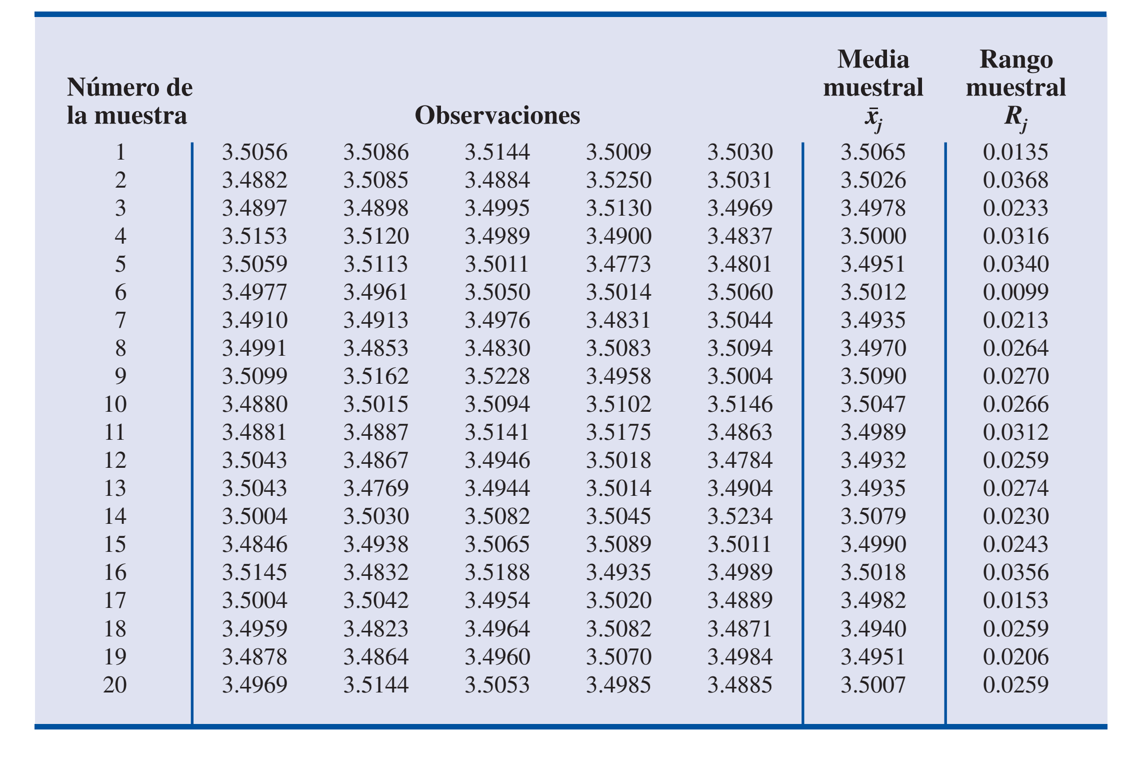
Con el ejemplo de la empresa KJW se mostró cómo elaborar una carta $\bar{x}$ cuando se conocen la media y la desviación estándar del proceso. Sin embargo, en la mayoría de los casos es necesario estimar la media y la desviación estándar del proceso a partir de muestras tomadas del proceso cuando éste se encuentre bajo control. Por ejemplo, durante 10 días en los que el proceso está bajo control, KJW puede tomar muestras aleatorias de 5 paquetes por las mañanas y 5 por las tardes, calcular la media y la desviación estándar de cada muestra, sacar después los promedios de las medias y de las desviaciones estándar y usarlos para elaborar cartas de control, tanto para la media como para la desviación estándar del proceso.

Es importante controlar tanto la media como la desviación estándar del proceso.

En la práctica, para monitorear la variabilidad del proceso suele emplearse el rango en lugar de la desviación estándar, ya que el rango es más fácil de calcular. El rango puede servir para obtener una buena estimación de la desviación estándar del proceso; de manera que, mediante algunos cálculos, puede emplearse para trazar los límites inferior y superior de las cartas $\bar{x}$. Para ilustrar esto, se tomará un ejemplo de la empresa Jensen Computer Supplies, Inc.

Jensen Computer Supplies (JCS) fabrica discos para computadora de 3.5 pulgadas de diámetro. Su proceso de producción acaba de ser ajustado, de manera que funciona bajo control. Su-

TABLA 20.2 DATOS DEL PROBLEMA DE JENSEN COMPUTER SUPPLIES



Número de la muestra	Observaciones					Media muestral $\bar{x}_j$	Rango muestral R_j
1	3.5056	3.5086	3.5144	3.5009	3.5030	3.5065	0.0135
2	3.4882	3.5085	3.4884	3.5250	3.5031	3.5026	0.0368
3	3.4897	3.4898	3.4995	3.5130	3.4969	3.4978	0.0233
4	3.5153	3.5120	3.4989	3.4900	3.4837	3.5000	0.0316
5	3.5059	3.5113	3.5011	3.4773	3.4801	3.4951	0.0340
6	3.4977	3.4961	3.5050	3.5014	3.5060	3.5012	0.0099
7	3.4910	3.4913	3.4976	3.4831	3.5044	3.4935	0.0213
8	3.4991	3.4853	3.4830	3.5083	3.5094	3.4970	0.0264
9	3.5099	3.5162	3.5228	3.4958	3.5004	3.5090	0.0270
10	3.4880	3.5015	3.5094	3.5102	3.5146	3.5047	0.0266
11	3.4881	3.4887	3.5141	3.5175	3.4863	3.4989	0.0312
12	3.5043	3.4867	3.4946	3.5018	3.4784	3.4932	0.0259
13	3.5043	3.4769	3.4944	3.5014	3.4904	3.4935	0.0274
14	3.5004	3.5030	3.5082	3.5045	3.5234	3.5079	0.0230
15	3.4846	3.4938	3.5065	3.5089	3.5011	3.4990	0.0243
16	3.5145	3.4832	3.5188	3.4935	3.4989	3.5018	0.0356
17	3.5004	3.5042	3.4954	3.5020	3.4889	3.4982	0.0153
18	3.4959	3.4823	3.4964	3.5082	3.4871	3.4940	0.0259
19	3.4878	3.4864	3.4960	3.5070	3.4984	3.4951	0.0206
20	3.4969	3.5144	3.5053	3.4985	3.4885	3.5007	0.0259



ponga que durante la primera hora de operación se toma una muestra aleatoria de cinco discos, durante la segunda hora de operación se toma otra muestra aleatoria de cinco discos y así sucesivamente, hasta que se tienen 20 muestras. En la tabla 20.2 se presentan los diámetros de las muestras así como la media $\bar{x}_j$ y el rango R_j de cada muestra.

La estimación de la media del proceso μ esta dada por la media muestral general.

MEDIA MUESTRAL GENERAL

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k} \quad \textbf{(20.4)}$$

donde

$$\bar{x}_j = \text{media de la muestra } j,\ j = 1, 2, \ldots, k$$
$$k = \text{número de muestras}$$

La media muestral general de los datos de JCS, que se presentan en la tabla 20.2, es $\bar{\bar{x}} = 3.4995$. Este valor será la línea central de la carta $\bar{x}$. El rango de cada muestra es simplemente la diferencia entre el valor mayor y el valor menor de cada muestra. El rango promedio de las k muestras se calcula como se indica a continuación.

RANGO PROMEDIO

$$\bar{R} = \frac{R_1 + R_2 + \cdots + R_k}{k} \quad \textbf{(20.5)}$$

donde

R_j = rango de la muestra j, $j = 1, 2, ..., k$
k = número de muestras

El rango promedio de los datos de JCS, que se presentan en la tabla 20.2, es $\bar{R} = 0.0253$.

En la sección anterior se mostró que los límites de control superior e inferior de la carta $\bar{x}$ son

$$\bar{x} \pm 3\frac{\sigma}{\sqrt{n}} \qquad \textbf{(20.6)}$$

La media muestral general $\bar{\bar{x}}$ se usa para estimar μ y los rangos muestrales se usan para obtener una estimación de σ.

Por tanto, para obtener los límites de control para la carta $\bar{x}$ es necesario estimar μ y σ, la media y la desviación estándar del proceso. Una estimación de μ es dada por $\bar{\bar{x}}$, y una estimación de σ se obtiene mediante los datos de los rangos.

Se puede demostrar que el promedio de los rangos dividido entre d_2, una constante que depende del tamaño n de la muestra, es una estimación de la desviación estándar σ del proceso. Es decir,

$$\text{Estimador de } \sigma = \frac{\bar{R}}{d_2} \qquad \textbf{(20.7)}$$

En el *Manual on Presentation of Data and Control Chart Analysis* de la *American Society for Testing and Materials* se publican los valores de d_2 que se muestran en la tabla 20.3. Por ejemplo, para $n = 5$, $d_2 = 2.326$ y la estimación de s es el rango promedio dividido entre 2.326. Si en la expresión (20.6) se sustituye s por $\bar{R}/d_2$, los límites de control de la carta $\bar{x}$ se pueden expresar como

$$\bar{\bar{x}} \pm 3\frac{\bar{R}/d_2}{\sqrt{n}} = \bar{\bar{x}} \pm \frac{3}{d_2\sqrt{n}}\bar{R} = \bar{\bar{x}} \pm A_2\bar{R} \qquad \textbf{(20.8)}$$

Observe que $A_2 = 3/(d_2\sqrt{n})$ es una constante que depende únicamente del tamaño de la muestra; los valores de A_2 también se encuentran en la tabla 20.3. Para $n = 5$, $A_2 = 0.577$; por tanto, los límites de control en la carta $\bar{x}$ serán

$$3.4995 \pm (0.577)(0.0253) = 3.4995 \pm 0.0146$$

Por tanto, UCL = 3.514 y LCL = 3.485.

En la figura 20.7 se muestra la carta $\bar{x}$ obtenida para el problema de Jensen Computer Supplies. Esta carta se obtuvo con los datos de la tabla 20.2 y la rutina para cartas de control de Minitab. La línea central se encuentra en el valor de la media muestral general $\bar{\bar{x}} = 3.4995$. El límite de control superior (UCL) es 3.514 y el límite de control inferior (LCL) es 3.485. En la carta $\bar{x}$ aparecen las 20 medias muestrales que se fueron graficando. Como todas las medias muestrales se encuentran dentro de los límites de control, se confirma que el proceso ha estado bajo control durante el periodo de muestreo.

Cartas *R*

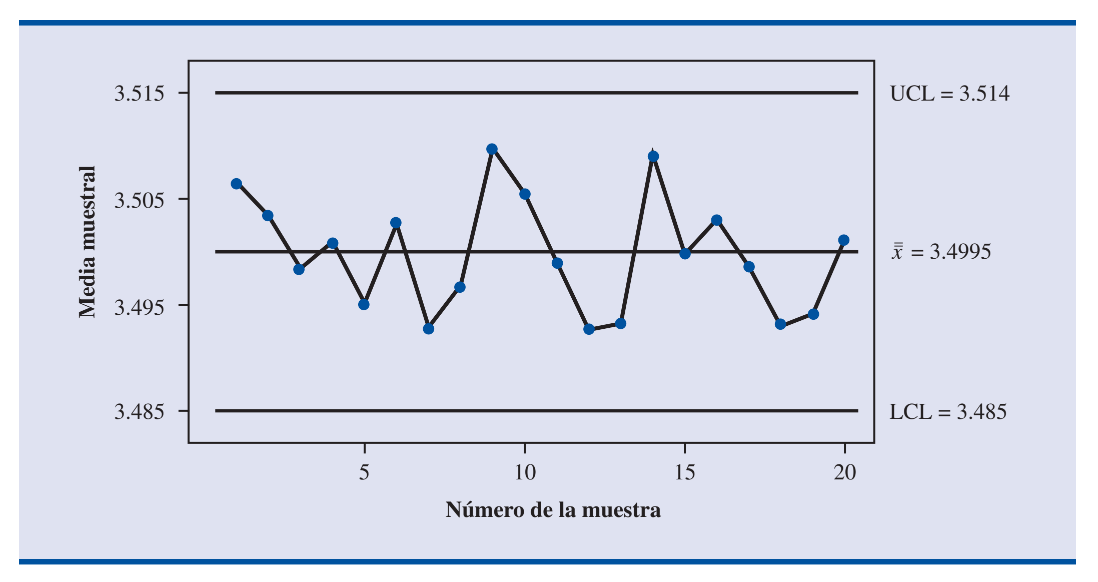
Ahora se estudiarán las cartas de rango (cartas R) que se emplean para controlar la variabilidad del proceso. Para elaborar una carta R es necesario considerar el rango de una muestra como una variable aleatoria con su media y desviación estándar propias. El rango promedio $\bar{R}$ proporciona

TABLA 20.3 FACTORES PARA LAS CARTAS $\bar{x}$ Y R

Observaciones en la muestra, n	d_2	A_2	d_3	D_3	D_4
2	1.128	1.880	0.853	0	3.267
3	1.693	1.023	0.888	0	2.574
4	2.059	0.729	0.880	0	2.282
5	2.326	0.577	0.864	0	2.114
6	2.534	0.483	0.848	0	2.004
7	2.704	0.419	0.833	0.076	1.924
8	2.847	0.373	0.820	0.136	1.864
9	2.970	0.337	0.808	0.184	1.816
10	3.078	0.308	0.797	0.223	1.777
11	3.173	0.285	0.787	0.256	1.744
12	3.258	0.266	0.778	0.283	1.717
13	3.336	0.249	0.770	0.307	1.693
14	3.407	0.235	0.763	0.328	1.672
15	3.472	0.223	0.756	0.347	1.653
16	3.532	0.212	0.750	0.363	1.637
17	3.588	0.203	0.744	0.378	1.622
18	3.640	0.194	0.739	0.391	1.608
19	3.689	0.187	0.734	0.403	1.597
20	3.735	0.180	0.729	0.415	1.585
21	3.778	0.173	0.724	0.425	1.575
22	3.819	0.167	0.720	0.434	1.566
23	3.858	0.162	0.716	0.443	1.557
24	3.895	0.157	0.712	0.451	1.548
25	3.931	0.153	0.708	0.459	1.541

Fuente: Adaptada de la tabla 27 de ASTM STP 15D, *ASTM Manual on Presentation of Data and Control Chart Analysis*. Copyright 1976 American Society for Testing and Materials, Filadelfia, PA. Impreso con autorización.

FIGURA 20.7 CARTA $\bar{x}$ PARA EL PROBLEMA DE JENSEN COMPUTER SUPPLIES



una estimación de la media de esta variable aleatoria. Además, se puede demostrar que una estimación de la desviación estándar del rango es

$$\hat{\sigma}_R = d_3 \frac{\bar{R}}{d_2} \quad \textbf{(20.9)}$$

donde d_2 y d_3 son constantes que dependen del tamaño de la muestra, y cuyos valores se encuentran también en la tabla 20.3. Por tanto, el UCL de la carta R está dado por

$$\bar{R} + 3\hat{\sigma}_R = \bar{R}\left(1 + 3\frac{d_3}{d_2}\right) \quad \textbf{(20.10)}$$

y el LCL es

$$\bar{R} - 3\hat{\sigma}_R = \bar{R}\left(1 - 3\frac{d_3}{d_2}\right) \quad \textbf{(20.11)}$$

Si se hace

$$D_4 = 1 + 3\frac{d_3}{d_2} \quad \textbf{(20.12)}$$

$$D_3 = 1 - 3\frac{d_3}{d_2} \quad \textbf{(20.13)}$$

Los límites de control de la carta R se expresan como

$$\text{UCL} = \bar{R}D_4 \quad \textbf{(20.14)}$$

$$\text{LCL} = \bar{R}D_3 \quad \textbf{(20.15)}$$

Los valores de D_3 y D_4 se dan también en la tabla 20.3. Observe que para $n = 5$, $D_3 = 0$ y $D_4 = 2.114$. Por tanto, como $\bar{R} = 0.0253$, los límites de control son

$$\text{UCL} = 0.0253(2.114) = 0.053$$

$$\text{LCL} = 0.0253(0) = 0$$

Si la carta R indica que el proceso está fuera de control, la carta $\bar{x}$ no deberá interpretarse hasta que la carta R indique que la variabilidad del proceso está bajo control.

En la figura 20.8 se muestra la carta R del problema de Jensen Computer Supplies. Esta carta se obtuvo con los datos de la tabla 20.2 y la rutina para cartas de control de Minitab. La línea central aparece en el valor de la media general de los 20 rangos muestrales, $\bar{R} = 0.253$. El UCL es 0.053 y el LCL es 0.000. En la carta R se observan los 20 rangos muestrales que se fueron graficando. Como los 20 rangos muestrales se encuentran dentro de los límites de control, se confirma que durante el periodo de muestreo el proceso estuvo bajo control.

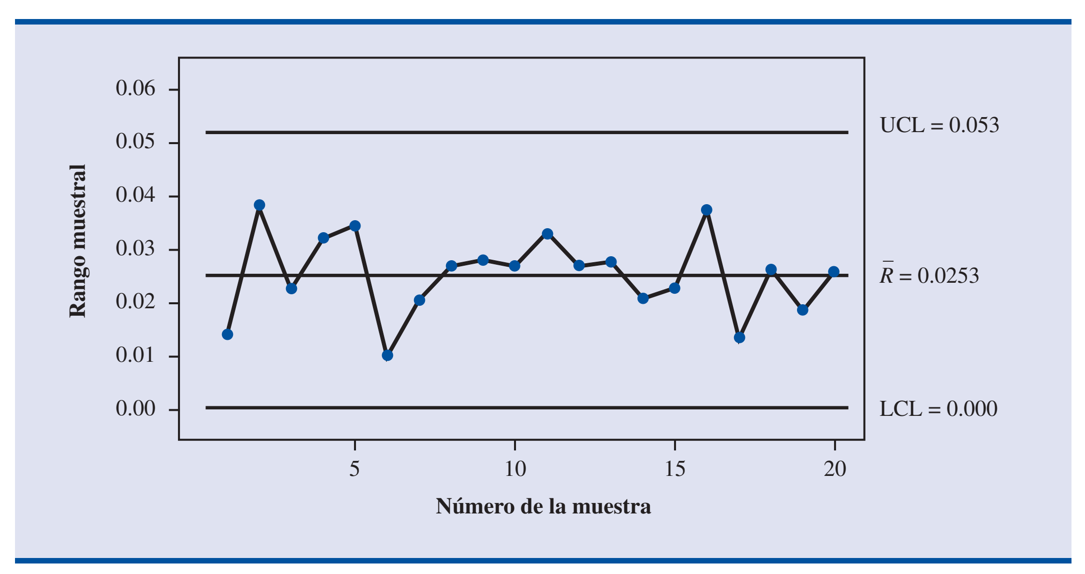
Cartas *p*

A las cartas de control que se basan en datos que indican la presencia de un defecto o un número de defectos se les llama cartas de control de atributos. Una carta p es una carta de control por atributos.

A continuación se considera el caso en que la calidad del producto se mide a partir de los artículos no defectuosos o los artículos defectuosos. La decisión de dejar que continúe el proceso de producción o que se ajuste se basa en $\bar{p}$, la proporción de artículos defectuosos encontrados en la muestra. A la carta de control que se usa para conocer la proporción de defectos se le llama carta p.

Para ilustrar la elaboración de una carta p, considere el uso de las máquinas automáticas para la clasificación de las cartas que se emplean en las oficinas de correo. Estas maquinas automáticas leen el código postal que aparece en el sobre y asignan la carta a la ruta de entrega correspondiente. Aun cuando la máquina esté funcionando en forma óptima, algunas de las cartas

FIGURA 20.8 CARTA *R* PARA EL PROBLEMA DE JENSEN COMPUTER SUPPLIES



no son asignadas correctamente. Suponga que cuando la máquina está en operación óptima o bajo control, 3% de las cartas no son asignadas correctamente. Entonces, p, la proporción de cartas no asignadas correctamente, con el proceso bajo control, es 0.03.

Para determinar la variación que puede esperarse en los valores de $\bar{p}$, cuando el proceso está bajo control, se usa la distribución muestral de $\bar{p}$, vista en el capítulo 7. Recuerde que el valor esperado, o la media, de $\bar{p}$, es p, la proporción de defectos cuando el proceso está bajo control. Si las muestras son de tamaño n, la fórmula para calcular la desviación estándar de $\bar{p}$, a la cual se le llama error estándar de la proporción, es

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{20.16}$$

En el capitulo 7 se vio también que la distribución muestral de $\bar{p}$, siempre que el tamaño de las muestras sea grande, es aproximable por medio de una distribución normal. El tamaño de la muestra puede considerarse grande siempre que se satisfagan las dos condiciones siguientes.

$$np \geq 5$$
$$n(1-p) \geq 5$$

En resumen, siempre que el tamaño de la muestra sea grande, la distribución muestral de $\bar{p}$ puede aproximarse mediante una distribución normal en que la media es p y la desviación estándar es $\sigma_{\bar{p}}$. Esta distribución se presenta en la figura 20.9.

Para establecer los límites de control en una carta p, se sigue el mismo procedimiento que se usó para establecer los límites de control en una carta $\bar{x}$. Es decir, los límites de control se establecen a 3 desviaciones estándar, o errores estándar, arriba y abajo de la proporción de defectos, cuando el proceso está bajo control. Por tanto, se tienen los siguientes límites de control.

LÍMITES DE CONTROL PARA UNA CARTA *p*

$$\text{UCL} = p + 3\sigma_{\bar{p}} \tag{20.17}$$
$$\text{LCL} = p - 3\sigma_{\bar{p}} \tag{20.18}$$

FIGURA 20.9 DISTRIBUCIÓN MUESTRAL DE $\bar{p}$

$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$

p

$E(\bar{p})$

$\bar{p}$

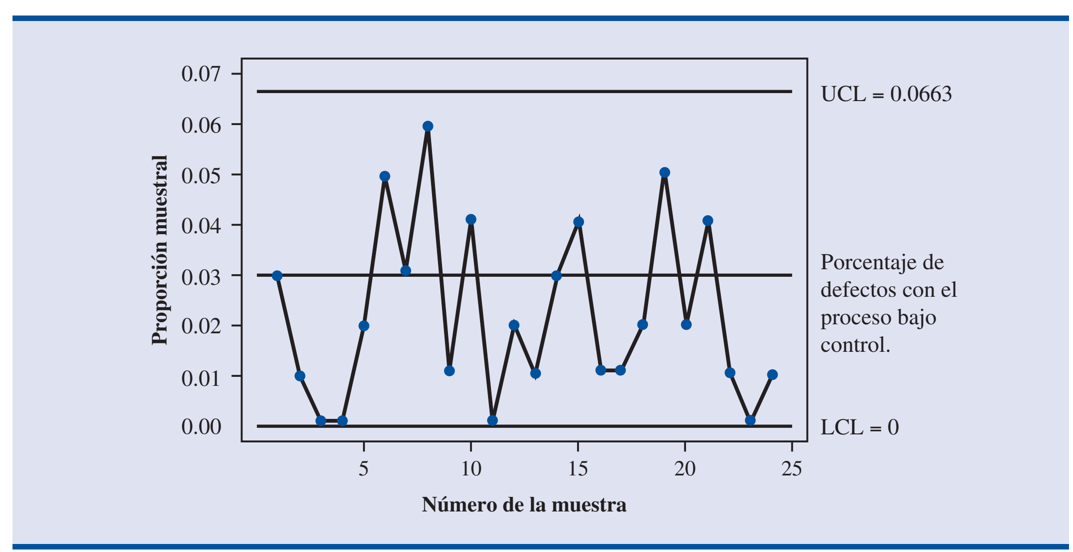
Como $p = 0.03$ y las muestras son de tamaño $n = 200$, al emplear la ecuación (20.16) se obtiene el error estándar, que es

$$\sigma_{\bar{p}} = \sqrt{\frac{0.03(1-0.03)}{200}} = 0.0121$$

Por tanto, los límites de control son, UCL = 0.03 + 3(0.0121) = 0.0663 y LCL = 0.03 − 3(0.0121) = −0.0063. Cuando con la ecuación (20.18) se obtiene un valor negativo para el LCL, se toma al cero como LCL para la carta de control.

La figura 20.10 es la carta de control del proceso de clasificación de las cartas. Los puntos muestreados son las proporciones muestrales de defectos halladas en las muestras de cartas tomadas del proceso. Todos los puntos se encuentran dentro de los límites de control, lo que indica que no hay evidencias para concluir que el proceso de clasificación de las cartas se encuentre fuera de control.

FIGURA 20.10 CARTA p PARA LA PROPORCIÓN DE DEFECTOS EN EL PROCESO DE CLASIFICACIÓN DE LAS CARTAS



Cuando no se conoce la proporción de defectos en un proceso que se encuentra bajo control, este valor se puede estimar a partir de datos muestrales. Suponga, por ejemplo, que de un proceso que se encuentra bajo control, se toman k muestras de tamaño n. Después se determina la proporción de defectos en cada muestra. Al considerar todos los datos obtenidos como una sola muestra grande, se calcula la proporción de artículos defectuosos en todos estos datos, y este valor se usa para estimar p, la proporción de artículos defectuosos observada con el proceso bajo control. Observe que dicha estimación de p también permite calcular el error estándar de la proporción, con lo que ya se pueden determinar los límites de control inferior y superior.

Cartas *np*

Una carta *np* es una carta de control que se usa para determinar el número de artículos defectuosos en una muestra. Aquí, n es el tamaño de la muestra y p es la probabilidad de observar un artículo defectuoso con el proceso bajo control. Siempre que el tamaño de la muestra sea grande, es decir, siempre que $np \geq 5$ y $n(1 - p) \geq 5$, la distribución del número de artículos defectuosos en una muestra de tamaño n puede aproximarse mediante una distribución normal con media np y desviación estándar $\sqrt{np(1 - p)}$. Entonces, en el ejemplo de la clasificación de las cartas, como $n = 200$ y $p = 0.03$, el número de artículos defectuosos observados en una muestra de 200 cartas se aproxima mediante una distribución normal con media de $200(0.03) = 6$ y la desviación estándar $\sqrt{200(0.03)(0.97)} = 2.4125$.

Los límites de control de una carta *np* se fijan a 3 desviaciones estándar arriba y abajo del número de artículos defectuosos esperado con el proceso bajo control. Por tanto, se tienen los límites de control siguientes

LÍMITES DE CONTROL EN UNA CARTA *np*

$$UCL = np + 3\sqrt{np(1 - p)} \tag{20.19}$$

$$LCL = np - 3\sqrt{np(1 - p)} \tag{20.20}$$

En el ejemplo de la clasificación de las cartas, como $p = 0.03$ y $n = 200$, los límites de control son $UCL = 6 + 3(2.4125) = 13.2375$ y $LCL = 6 - 3(2.4125) = -1.2375$. Cuando el LCL es negativo, como LCL de la carta de control se toma igual a cero. Por tanto, si el número de cartas que no se asigna a la ruta correcta es mayor que 13, se concluye que el proceso está fuera de control.

La información que proporciona una carta *np* es similar a la información que proporciona una carta p; la única diferencia es que la carta *np* es la gráfica del número de artículos defectuosos observados, mientras que una carta p es la gráfica de la proporción de artículos defectuosos observados. De manera que si se concluye que un proceso está fuera de control, con base en una carta p, también se concluirá apoyándose en una carta *np* que el proceso está fuera de control.

Interpretación de las cartas de control

La ubicación y el patrón que siguen los puntos en una carta de control permiten determinar, con una pequeña probabilidad de error, si un proceso se encuentra estadísticamente bajo control. Una primera indicación de que un proceso pueda estar fuera de control es que uno de los puntos de los datos se encuentre fuera de los límites de control, como ocurre con el punto 5 de la figura 20.6. Hallar uno de estos puntos es evidencia estadística de que el proceso se encuentra fuera de control. En tales casos deberán tomarse medidas correctivas tan pronto como sea posible.

Además de la presencia de puntos fuera de los límites de control, hay ciertos patrones de los puntos, dentro de los límites de control, que pueden ser señales que adviertan de problemas de control de calidad. Por ejemplo, suponga que todos los puntos están dentro de los límites de con-

trol, pero que muchos de ellos se encuentran de un mismo lado de la línea central. Este patrón puede indicar que existe algún problema en el equipo, que ha habido una variación en los materiales o que alguna otra causa asignable ha ocasionado una variación de la calidad. Será necesario hacer una cuidadosa investigación del proceso para determinar si la calidad ha variado.

Aun cuando todos los puntos se encuentren entre los límites de control inferior y superior, el proceso puede encontrarse fuera de control. Tendencias en los puntos muestrales o secuencias inusualmente largas de puntos que se encuentren sobre o bajo la línea central, también pueden indicar una situación fuera de control.

Otro patrón a observar en una carta de control es si existe un desplazamiento gradual o una tendencia, a lo largo del tiempo. Por ejemplo, debido al desgaste de las herramientas, las dimensiones en la fabricación de una pieza pueden desviarse gradualmente de las medidas establecidas. Variaciones graduales de la temperatura o de la humedad, el deterioro gradual del equipo, la acumulación de suciedad o el cansancio de un operador pueden ocasionar la aparición de tendencias en las cartas de control. Seis o siete puntos consecutivos que muestren una tendencia deberán ser causa de preocupación, aun cuando todos estos puntos se encuentren dentro de los límites de control. Siempre que se observen tales patrones, deberá revisarse el proceso para determinar si hay algún cambio o desplazamiento en la calidad. Puede que sea necesario tomar medidas correctivas para que el proceso vuelva a estar bajo control.

NOTAS Y COMENTARIOS

1. Como en las cartas $\bar{x}$ los límites de control dependen del valor del rango promedio, estos límites no tienen mucho significado a menos que la variabilidad del proceso se encuentre bajo control. En la práctica, la carta R se elabora antes que la carta $\bar{x}$; si la carta R indica que la variabilidad del proceso está bajo control, entonces se elabora la carta $\bar{x}$. Con la opción de Minitab Xbar-R se obtienen, simultáneamente, la carta $\bar{x}$ y la carta R. En el apéndice 20.1 se describen los pasos a seguir con Minitab para obtener estas cartas.

2. Las cartas *np* se usan para monitorear un proceso en términos del número de defectos. El nivel de calidad Seis sigma de Motorola tiene como objetivo producir no más de 3.4 defectos por millón de operaciones; este objetivo implica $p = 0.0000034$.

Ejercicios

Métodos

1. En un proceso que está bajo control, la media es $\mu = 12.5$ y la desviación estándar es $\sigma = 0.8$.
   a. Elabore la carta de control $\bar{x}$ para este proceso; el tamaño de las muestras es 4.
   b. Repita el inciso a con muestras de tamaño 8 y 16.
   c. ¿Qué pasa con los límites de la carta de control a medida que aumenta el tamaño de la muestra? Analice por qué esto es razonable.
2. En un proceso que está bajo control se toman 25 muestras de tamaño 5. La suma de todos los datos recolectados fue 677.5 libras.
   a. Dé una estimación de la media del proceso (en términos de libras por unidad) cuando el proceso está bajo control.
   b. Elabore la carta de control $\bar{x}$ para este proceso, considere que las muestras son de tamaño 5. Suponga que cuando el proceso está bajo control la desviación estándar del proceso es 0.5 y la media del proceso es la estimación que se obtuvo en el inciso a.
3. Mientras un proceso funcionaba satisfactoriamente se tomaron 25 muestras de 100 artículos cada una. En estas 25 muestras se encontraron 135 artículos defectuosos.
   a. Dé una estimación de la proporción de defectos que hay cuando el proceso está bajo control.
   b. ¿Cuál es el error estándar de la proporción si para el control estadístico del proceso se usan muestras de tamaño 100?
   c. Calcule los límites de control inferior y superior para la carta de control.



4. En un proceso del que se toman 20 muestras de tamaño 8 cada una, $\bar{\bar{x}} = 28.5$ y $\bar{R} = 1.6$. Calcule los límites de control inferior y superior de las cartas $\bar{x}$ y R del proceso.

Aplicaciones

5. Para medir los resultados de un proceso de producción se emplea la temperatura. Cuando el proceso está bajo control, la media del proceso es $\mu = 128.5$ y la desviación estándar es $\sigma = 0.4$.
   a. Construya la carta $\bar{x}$ de este proceso con muestras empleadas de tamaño 6.
   b. Si los datos que se obtienen de una muestra son los siguientes, ¿el proceso está bajo control?

   128.8 128.2 129.1 128.7 128.4 129.2

   c. Si los datos que se obtienen de una muestra son los siguientes, ¿está bajo control el proceso?

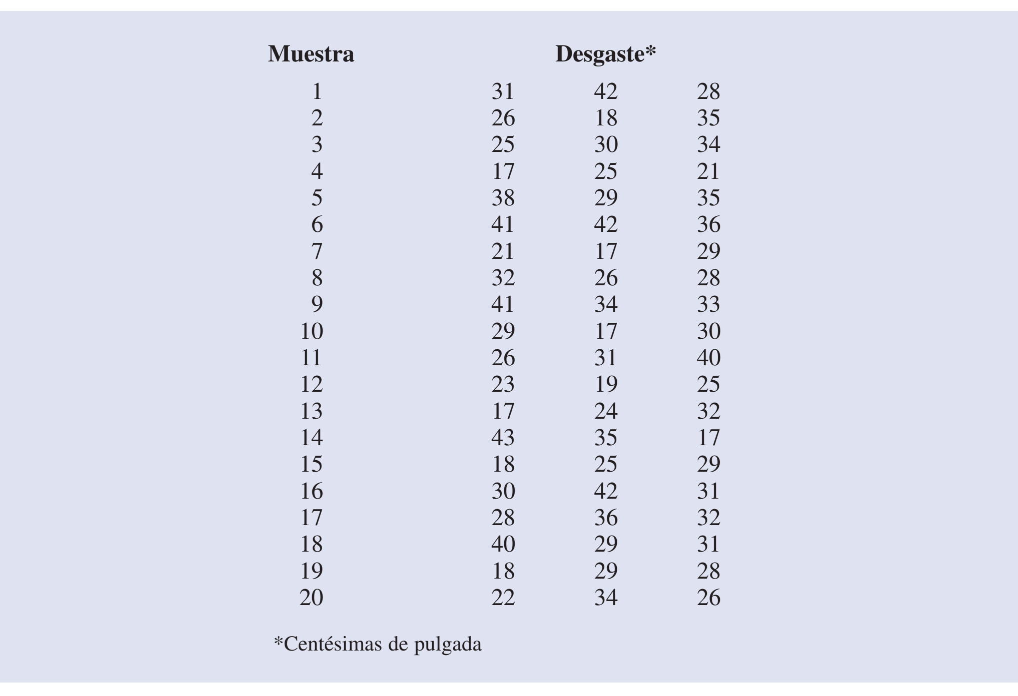
   129.3 128.7 128.6 129.2 129.5 129.0

6. En un proceso de control de calidad se vigila el peso por paquete de detergente envasado. Los límites de control estipulados son UCL = 20.12 onzas y LCL = 19.90 onzas. En este proceso de muestreo e inspección se emplean muestras de tamaño 5. Dé la media y la desviación estándar del proceso.

7. La empresa Goodman Tire and Rubber hace pruebas periódicas a sus neumáticos para determinar su desgaste. Para estudiar y controlar el proceso de fabricación, durante varios días y de los diferentes turnos se tomaron 20 muestras, cada una de tres neumáticos radiales. Estos datos se presentan a continuación. Si estos datos se obtuvieron cuando se creía que el proceso estaba bajo control, elabore las cartas R y $\bar{x}$.



Tires



Muestra		Desgaste*	
1	31	42	28
2	26	18	35
3	25	30	34
4	17	25	21
5	38	29	35
6	41	42	36
7	21	17	29
8	32	26	28
9	41	34	33
10	29	17	30
11	26	31	40
12	23	19	25
13	17	24	32
14	43	35	17
15	18	25	29
16	30	42	31
17	28	36	32
18	40	29	31
19	18	29	28
20	22	34	26

*Centésimas de pulgada

8. A lo largo varias semanas de funcionamiento normal, o bajo control, se tomaron 20 muestras,cada una de 150 paquetes de cuerdas sintéticas para raquetas de tenis, y se probó su resistencia a la rotura. De los 3 000 paquetes probados, 141 no satisficieron las especificaciones del fabricante.
   a. Dé una estimación de la proporción de defectos en el proceso cuando el proceso se encuentra bajo control.
   b. Calcule los límites superior e inferior de la carta *p*.
   c. De acuerdo con los resultados del inciso b, qué conclusiones puede obtener acerca del proceso si al probar una nueva muestra de 150 paquetes se encuentran 12 defectuosos. ¿Existen algunas causas asignables en esta situación?
   d. Calcule los límites inferior y superior de la carta *np*.

e. Responda el inciso c, use los resultados del inciso d.
f. ¿Qué carta de control convendrá para esta situación? Explique.

9. Un proveedor de la industria automotriz produce pistones para varios modelos de automóviles. Cuando el proceso se encuentra bajo control se toman 20 muestras, cada una de 200 pistones. A continuación se presenta el número de pistones defectuosos hallados en cada muestra.

8	10	6	4	5	7	8	12	8	15
14	10	10	7	5	8	6	10	4	8

a. Dé una estimación de la proporción de defectos en el proceso de fabricación cuando el proceso está bajo control.
b. Elabore la carta *p* de este proceso, si cada muestra tiene 200 pistones.
c. De acuerdo con los resultados del inciso b, ¿cuál es la conclusión si en una muestra de 200 pistones se encuentran 20 defectuosos?
d. Calcule los límites superior e inferior de la carta *np*.
e. Responda el inciso c, use los resultados del inciso d.



Muestreo de aceptación

En el muestreo de aceptación los objetos de interés pueden ser entregas de materias primas o de partes o de bienes terminados. Suponga que desea aceptar o rechazar un conjunto de artículos a partir de determinadas características de calidad. En la terminología del control de calidad, al conjunto de artículos se le conoce como **lote**, y el **muestreo de aceptación** es un método estadístico que permite basar la decisión de aceptar o rechazar el lote en la inspección de una muestra de los artículos del lote.

En la figura 20.11 se muestran los pasos que, en general, se siguen en el muestreo de aceptación. Después de recibir un lote se toma una muestra de artículos para su inspección. Los resultados de la inspección se comparan con las características de calidad especificadas. Si se satisfacen las características de calidad, el lote se acepta y se envía a producción o al cliente. Si el lote se rechaza, los directivos tendrán que decidir cómo se dispone del lote. En algunos casos la decisión puede ser quedarse con el lote y eliminar los artículos que no son aceptables. En otros casos se devuelve el lote al proveedor, a su cargo; el trabajo extra y los costos que se le cargan al proveedor puede que hagan que el proveedor mejore la calidad de sus productos. Por último, si el lote rechazado consta de bienes terminados, estos bienes tendrán que desecharse o adaptarse a los estándares de calidad aceptables.

Las ventajas del muestreo de aceptación sobre la inspección 100% son:
1. Mucho menos costoso
2. Menos daño al producto debido a menor manipulación y menos pruebas
3. Se requieren menos inspectores
4. Es la única posibilidad cuando hay que emplear una prueba destructiva

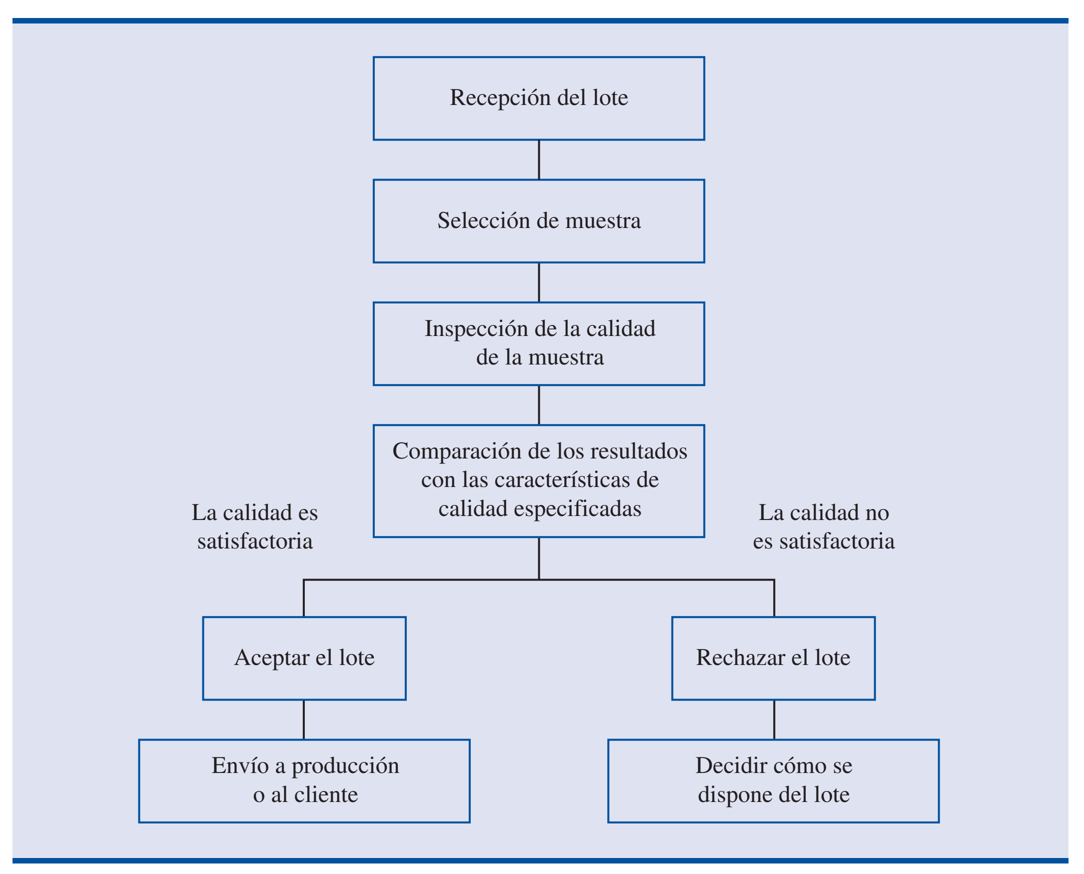
El procedimiento estadístico del muestreo de aceptación se basa en la metodología de la prueba de hipótesis presentada en el capítulo 9. Las hipótesis nula y alternativa son las siguientes.

$$H_0\text{: La calidad del lote es buena}$$
$$H_a\text{: La calidad del lote es mala}$$

En la tabla 20.4 se muestran los resultados de la prueba de hipótesis. Observe que las decisiones correctas son aceptar un lote de buena calidad y rechazar un lote de mala calidad. Sin embargo, como en las demás pruebas de hipótesis, hay que tener cuidado de no cometer un error tipo I (rechazar un lote de buena calidad) o un error tipo II (aceptar un lote de mala calidad).

La probabilidad de cometer un error tipo I representa un riesgo para el productor del lote y se conoce como el **riesgo del productor**. Por ejemplo, un riesgo del productor de 0.05 significa que existe 5% de posibilidad de que un lote de buena calidad sea rechazado erróneamente. Por otro lado, la probabilidad de cometer un error tipo II, representa un riesgo para el consumidor del lote y se conoce como **riesgo del consumidor**. Por ejemplo, un riesgo del consumidor de 0.10 significa que existe 10% de posibilidad de que un lote de mala calidad sea erróneamente aceptado y usado para la producción o para surtir un pedido al cliente. La persona que elabora el procedimiento de aceptación de muestras puede controlar los valores que determinan el riesgo del productor y el riesgo del consumidor. Para ilustrar cómo se determinan estos valores de riesgo se tomará un problema que se le presentó a la empresa KALI, Inc.

FIGURA 20.11 PROCEDIMIENTO DE MUESTREO DE ACEPTACIÓN



KALI, Inc., un ejemplo de muestreo de aceptación

KALI, Inc., fabrica aparatos para el hogar que se venden bajo diferentes marcas. KALI no fabrica todos los componentes que usa en sus productos, varios de ellos los compra. Por ejemplo, uno de los componentes que compra KALI es uno que emplea en sus equipos para aire acondicionado, se trata de un protector de sobrecarga, un dispositivo que desconecta la compresora cuando ésta se sobrecalienta. Si el protector de sobrecarga no funciona adecuadamente, la compresora puede sufrir un grave daño, por lo que KALI pone mucha atención a la calidad de estos protectores de sobrecarga. Una manera de asegurarse de la calidad adecuada de estos dispositivos es probar cada uno de ellos con un método que se conoce como inspección 100%. Sin embargo,

TABLA 20.4 RESULTADOS DEL MUESTREO DE ACEPTACIÓN

		Estado de lote	
		H_0 verdadera **La calidad de lote es buena**	**H_0 falsa** **La calidad de lote es mala**
Decisión	**Aceptar el lote**	Desición correcta	Error tipo II (aceptar un lote de mala calidad)
	Rechazar el lote	Error tipo I (rechazar un lote de buena calidad)	Decisión correcta

para esto, el protector de sobrecarga debería someterse a pruebas que cuestan tiempo y dinero, y KALI encuentra justificable probar cada protector de carga que compra.

Lo que hace KALI es emplear un plan de muestreo de aceptación para monitorear la calidad de los protectores de sobrecarga. Para el plan de muestreo de aceptación se requiere que los inspectores de control de calidad de KALI elijan y prueben una muestra de protectores de sobrecarga de cada pedido que reciben. Si en la muestra sólo se encuentran unas cuantas unidades defectuosas, es probable que la calidad del lote sea buena y entonces se aceptará. Pero, si en la muestra se encuentra un gran número de unidades defectuosas, es probable que la calidad del lote sea mala y deberá rechazarse.

Un plan para un muestreo de aceptación consta de un tamaño n de muestra y un criterio de aceptación c. El **criterio de aceptación** es el número máximo de artículos defectuosos que se puede encontrar en una muestra para que el lote se considere aceptable. Por ejemplo, en el caso de KALI, suponga que de cada lote o pedido que se reciba se tomará una muestra de 15 artículos. Además, el gerente de control de calidad decide que los lotes sólo se pueden aceptar si no se encuentra ningún artículo defectuoso. En tal caso el plan de muestreo de aceptación establecido por el gerente de control es $n = 15$ y $c = 0$.

Este plan de muestreo de aceptación será fácil de realizar para los inspectores de control de calidad: simplemente tienen que tomar una muestra de 15 artículos, realizar la prueba y llegar a una conclusión con base en la siguiente regla de decisión.

- *Aceptar el lote* si encuentran cero artículos defectuosos
- *Rechazar el lote* si encuentran uno o más artículos defectuosos.

Antes de poner en marcha este plan de muestreo de aceptación, el gerente de control de calidad quiere evaluar los riesgos o posibles errores que se pueden tener en este plan. Para poner el plan en marcha es necesario que tanto el riesgo del productor (error tipo I) como el riesgo del consumidor (error tipo II) estén controlados en niveles razonables.

Cálculo de la probabilidad de aceptar un lote

La clave para analizar tanto el riesgo del productor como el riesgo del consumidor es un análisis del tipo "y qué pasa si". Es decir, se supone que el lote tiene un determinado número de artículos defectuosos y se calcula la probabilidad de aceptar el lote con un determinado plan de muestreo. Al variar el porcentaje de artículos defectuosos que se está dispuesto a aceptar, es posible examinar el efecto del plan de muestreo sobre los dos tipos de riesgos.

Para empezar, si se ha recibido un pedido grande de protectores de sobrecarga y 5% de los protectores de sobrecarga de este pedido están defectuosos. ¿Cuál es la probabilidad de que con el plan de aceptación $n = 15$, $c = 0$ se acepte un lote en que 5% de los artículos está defectuoso? Como cada protector de sobrecarga que se prueba estará defectuoso o no estará defectuoso y como el lote es grande, el número de artículos defectuosos en una muestra de tamaño 15 tendrá una *distribución binomial*. A continuación se presenta la función de distribución binomial que se presentó en el capítulo 5.

FUNCIÓN DE ACEPTACIÓN BINOMIAL PARA EL MUESTREO DE ACEPTACIÓN

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)} \qquad \textbf{(20.21)}$$

donde

n = tamaño de la muestra
p = proporción de artículos defectuosos en el lote
x = número de artículos defectuosos en la muestra
$f(x)$ = probabilidad de que haya x artículos defectuosos en la muestra

En el plan de muestreo de aceptación de KALI, $n = 15$; por tanto, con un lote en el que hay 5% de defectos ($p = 0.05$), se tiene

$$f(x) = \frac{15!}{x!(15 - x)!}(0.05)^x(1 - 0.05)^{(15-x)} \tag{20.22}$$

Mediante la ecuación (20.22) con $f(0)$ se obtendrá la probabilidad de que 0 de los protectores de sobrecarga estén defectuosos y de que el lote sea aceptado. Al usar la ecuación (20.22) recuerde que $0! = 1$. El cálculo de la probabilidad $f(0)$ es

$$f(0) = \frac{15!}{0!(15 - 0)!}(0.05)^0(1 - 0.05)^{(15-0)}$$
$$= \frac{15!}{0!(15)!}(0.05)^0(0.95)^{15} = (0.95)^{15} = 0.4633$$

Ahora ya sabe que con el plan de muestreo $n = 15$ y $c = 0$ la probabilidad de aceptar un lote con 5% de defectos es 0.4633. Por tanto la probabilidad correspondiente de rechazar un lote con 5% de defectos es $1 - 0.4633 = 0.5367$.

Las probabilidades binomiales también se calculan usando Excel o Minitab.

El uso de las tablas de probabilidad binomial (véase tabla 5 del apéndice B) facilita los cálculos para determinar la probabilidad de aceptar un lote. En la tabla 20.5 se presentan algunas probabilidades binomiales para $n = 15$ y $n = 20$. Con esta tabla se puede determinar que si el lote

TABLA 20.5 ALGUNAS PROBABILIDADES BINOMIALES PARA MUESTRAS DE TAMAÑO 15 Y 20

		p								
n	*x*	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**
15	0	0.8601	0.7386	0.6333	0.5421	0.4633	0.2059	0.0874	0.0352	0.0134
	1	0.1303	0.2261	0.2938	0.3388	0.3658	0.3432	0.2312	0.1319	0.0668
	2	0.0092	0.0323	0.0636	0.0988	0.1348	0.2669	0.2856	0.2309	0.1559
	3	0.0004	0.0029	0.0085	0.0178	0.0307	0.1285	0.2184	0.2501	0.2252
	4	0.0000	0.0002	0.0008	0.0022	0.0049	0.0428	0.1156	0.1876	0.2252
	5	0.0000	0.0000	0.0001	0.0002	0.0006	0.0105	0.0449	0.1032	0.1651
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0019	0.0132	0.0430	0.0917
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0030	0.0138	0.0393
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0035	0.0131
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0007	0.0034
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0007
20	0	0.8179	0.6676	0.5438	0.4420	0.3585	0.1216	0.0388	0.0115	0.0032
	1	0.1652	0.2725	0.3364	0.3683	0.3774	0.2702	0.1368	0.0576	0.0211
	2	0.0159	0.0528	0.0988	0.1458	0.1887	0.2852	0.2293	0.1369	0.0669
	3	0.0010	0.0065	0.0183	0.0364	0.0596	0.1901	0.2428	0.2054	0.1339
	4	0.0000	0.0006	0.0024	0.0065	0.0133	0.0898	0.1821	0.2182	0.1897
	5	0.0000	0.0000	0.0002	0.0009	0.0022	0.0319	0.1028	0.1746	0.2023
	6	0.0000	0.0000	0.0000	0.0001	0.0003	0.0089	0.0454	0.1091	0.1686
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0020	0.0160	0.0545	0.1124
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0046	0.0222	0.0609
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0011	0.0074	0.0271
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0020	0.0099
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0030
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0008

TABLA 20.6 PROBABILIDADES DE ACEPTAR EL LOTE DEL PROBLEMA DE KALI CON $n = 15$ Y $c = 0$

Porcentaje de defectos en un lote	Probabilidad de aceptar el lote
1	0.8601
2	0.7386
3	0.6333
4	0.5421
5	0.4633
10	0.2059
15	0.0874
20	0.0352
25	0.0134

contiene 10% de artículos defectuosos la probabilidad de aceptarlo con el plan de muestreo $n = 15$ y $c = 0$ es 0.2059. En la tabla 20.6 se presentan las probabilidades de que el plan de muestreo $n = 15$ y $c = 0$ lleve a la aceptación de un lote con 1%, 2%, 3%, … artículos defectuosos.

Con las probabilidades de la tabla 20.6, se puede trazar la gráfica de la probabilidad de aceptar un lote frente al porcentaje de defectos en el lote como se muestra en la figura 20.12. A esta gráfica o a esta curva se le conoce como **curva característica de operación (curva CO)** del plan de muestreo de aceptación $n = 15$ y $c = 0$.

Quizá se deban considerar otros planes de muestreo, planes con otros tamaños de muestra n o con otros criterios de aceptación c. Considere primero el caso en que el tamaño de la muestra aún es $n = 15$, pero el criterio de aceptación aumenta de $c = 0$ a $c = 1$. Es decir, ahora se aceptará el lote si en la muestra se encuentran cero o un componente defectuoso. En un lote con 5% de artículos defectuosos ($p = 0.05$), en la tabla 20.5 se encuentra que para $n = 15$ y $p = 0.05$,

FIGURA 20.12 CURVA CARACTERÍSTICA DE OPERACIÓN PARA EL PLAN DE MUESTREO DE ACEPTACIÓN $N = 15$, $C = 0$

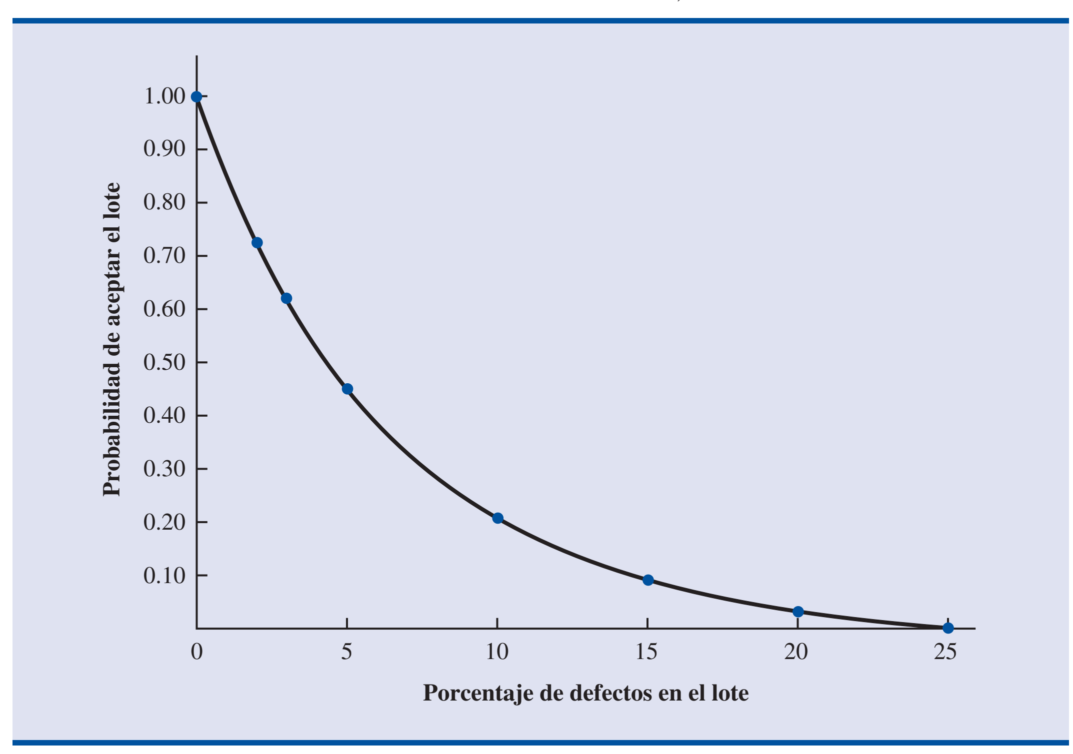
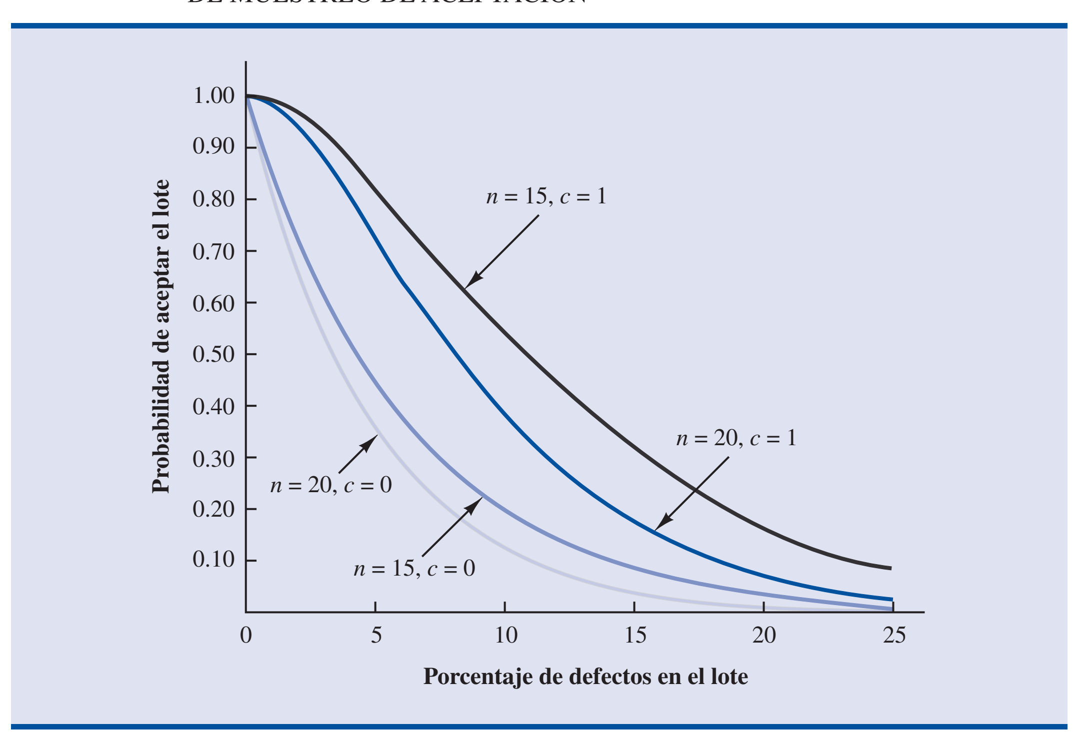


FIGURA 20.13 CURVAS CARACTERÍSTICAS DE OPERACIÓN DE CUATRO PLANES DE MUESTREO DE ACEPTACIÓN



$f(0) = 0.4633$ y $f(1) = 0.3658$. Por tanto, la probabilidad de que el plan $n = 15$, $c = 1$, lleve a la aceptación de un lote con 5% de defectos es $0.4633 + 0.3658 = 0.8291$.

Al continuar con estos cálculos se obtiene la figura 20.13, en la que se presentan las curvas características de operación de cuatro planes de muestreo de aceptación para el caso de KALI. Se consideraron muestras de tamaños 15 y 20. Observe que sea cual sea la proporción de defectos en un lote, con el plan de muestreo $n = 15$, $c = 1$ se tiene la mayor probabilidad de aceptar el lote. Con el plan $n = 20$, $c = 0$ se tienen la menor probabilidad de aceptar el lote. Pero con este plan se tiene también la mayor probabilidad de rechazar el lote.

Selección de un plan de muestreo de aceptación

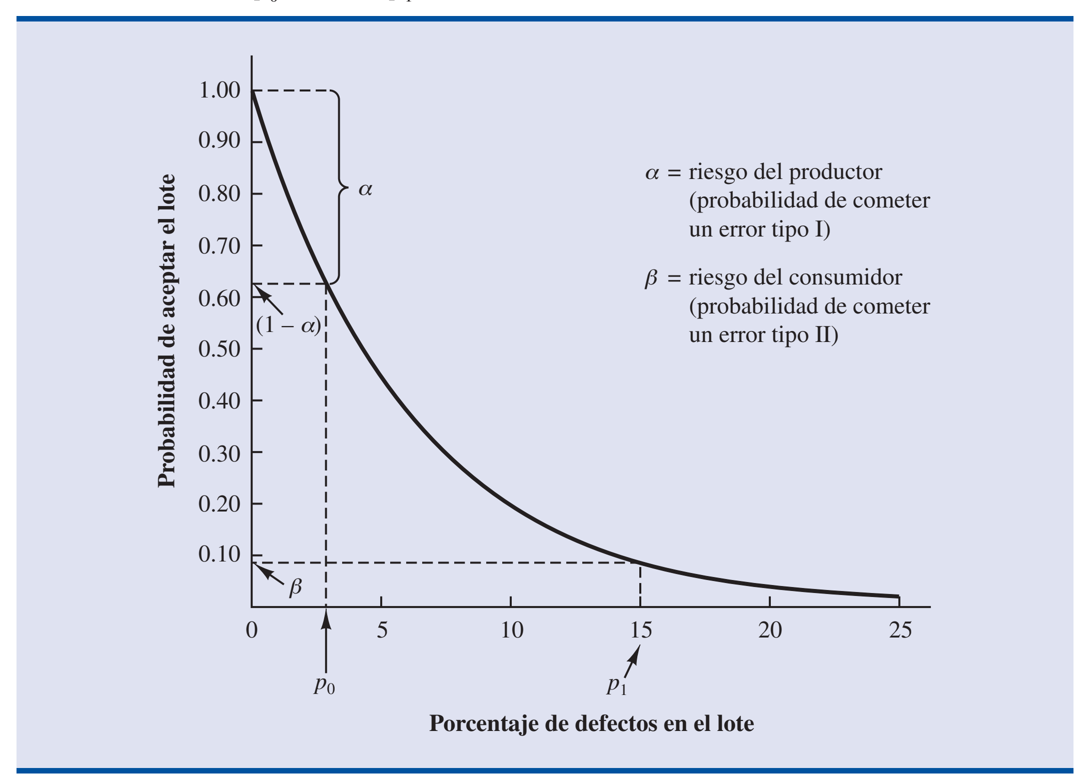
Ahora que ya se sabe usar la distribución binomial para calcular la probabilidad de aceptar un lote con una determinada proporción de defectos, ya es posible elegir los valores de n y c que determinen el plan de muestreo de aceptación deseado para el caso en estudio. Para obtener el plan, los directivos tendrán que especificar dos valores para la proporción de defectos en el lote. Un valor, denotado p_0, que se usa para determinar el riesgo del productor y otro valor, denotado p_1, que se usa para determinar el riesgo del consumidor.

Se usará la notación siguiente:

α = riesgo del productor; probabilidad de rechazar el lote con una proporción p_0 artículos defectuosos
β = riesgo del consumidor; probabilidad de aceptar el lote con una proporción p_1 artículos defectuosos

Suponga que en el caso de KALI, los directivos especifican $p_0 = 0.03$ y $p_1 = 0.15$. En la figura 20.14, en la curva CO para $n = 15$, $c = 0$, se observa que con $p_0 = 0.03$ el riesgo del productor es, aproximadamente, $1 - 0.63 = 0.37$ y con $p_1 = 0.15$ el riesgo del consumidor es, aproximadamente, 0.09. Por tanto, si los directivos están dispuestos a tolerar, tanto una probabilidad de

FIGURA 20.14 CURVA CARACTERÍSTICA DE OPERACIÓN PARA $n = 15$, $c = 0$ CON $p_0 = 0.03$ Y $p_1 = 0.15$



0.37 de rechazar un lote con 3% de artículos defectuosos (riesgo del productor) como una probabilidad de 0.09 de aceptar un lote con 15% de artículos defectuosos (riesgo del consumidor), entonces el plan de muestreo de aceptación $n = 15$, $c = 0$ será aceptable.

Pero, si los directivos desean que el riesgo del productor sea $\alpha = 0.10$ y que el riesgo del consumidor sea $\beta = 0.20$. Como se ve, con el plan de muestreo $n = 15$, $c = 0$, el riesgo del consumidor es mejor de lo deseado, pero el riesgo del productor es demasiado grande y no se puede aceptar. El que $\alpha = 0.37$ significa que 37% de los lotes se rechazarán erróneamente cuando tengan sólo 3% de artículos defectuosos. El riesgo del productor es demasiado alto y se tendrá que considerar otro plan de muestreo de aceptación.

En el ejercicio 13, que se encuentra al final de esta sección, se pedirá calcular el riesgo del productor y el riesgo del consumidor para el plan de muestreo $n = 20$, $c = 1$.

En la figura 20.13 se ve que para $p_0 = 0.3$, $\alpha = 0.10$, $p_1 = 0.15$ y $\beta = 0.20$, el plan de aceptación $n = 20$ y $c = 1$ está más cerca de los requerimientos para los riesgos del productor y del consumidor.

Como se muestra en esta sección, se necesitarán varios cálculos y diversas curvas de operación para determinar el plan de muestreo con los riesgos deseados para el productor y para el consumidor. Por fortuna existen tablas de planes de muestreo. Por ejemplo, en la American Military Standard Table, MIL-STD-105D, se encuentra útil información para el diseño de planes de muestreo de aceptación. El uso de estas tablas se describe en textos más avanzados sobre el control de calidad, como los citados en la bibliografía. En los textos más avanzados se estudia también la importancia del costo del muestreo al determinar el plan de muestreo óptimo.

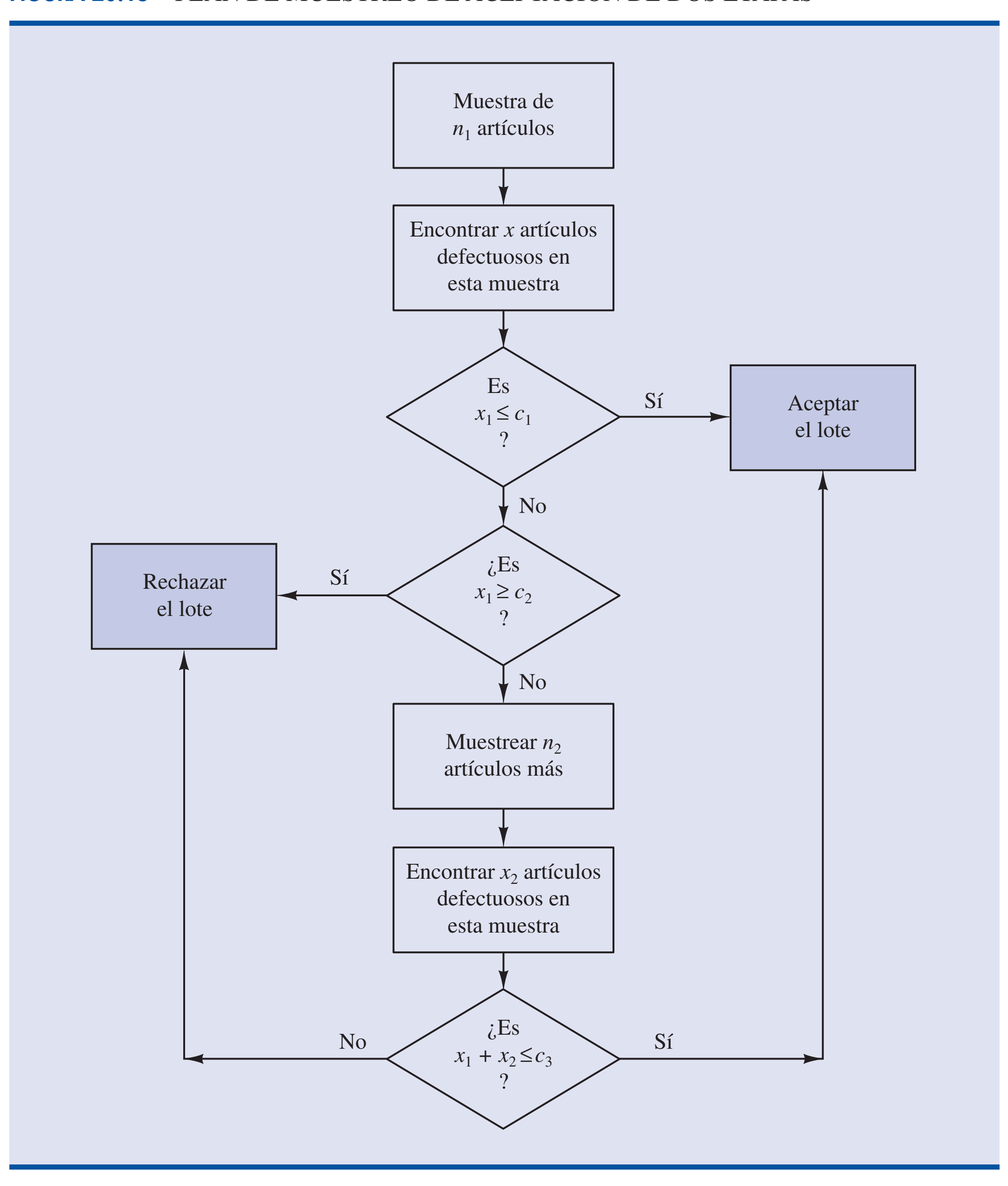
Planes de muestreo múltiple

El procedimiento de muestreo de aceptación presentado para el caso de KALI es un plan *sencillo de muestreo.* Se le llama plan sencillo de muestreo porque sólo se usa una muestra o un escenario. Una vez determinado el número de defectos en la muestra, hay que tomar la decisión de

aceptar o rechazar el lote. Una alternativa al plan sencillo de muestreo es el plan de **muestreo múltiple**, en el que se usan dos o más etapas de muestreo. En cada etapa hay que decidirse entre tres posibilidades: dejar de muestrear y aceptar el lote, dejar de muestrear y rechazar el lote o continuar muestreando. Aun cuando son más complejos, los planes de muestreo múltiples suelen dar como resultado tamaños de muestra más pequeños que los planes de muestreo sencillos con las mismas probabilidades a y b.

En la figura 20.15 se muestra la lógica del plan de las dos etapas, o de la doble muestra. Al inicio se toma una muestra de n_1 artículos. Si el número de elementos defectuosos x_1 es menor o igual a c_1, se acepta el lote. Si x_1 es mayor o igual a c_2, se rechaza el lote. Si x_1 se encuentra entre c_1 y c_2 ($c_1 < x_1 < c_2$), se toma una segunda muestra de n_2 artículos. Se determina la suma de los artículos defectuosos en la primer (x_1) y segunda (x_2) muestras. Si $x_1 + x_2 \leq c_3$ se acepta el lote, si no es así se rechaza el lote. El plan de doble muestra es más complicado, ya que los ta-

FIGURA 20.15 PLAN DE MUESTREO DE ACEPTACIÓN DE DOS ETAPAS



maños muestrales n_1 y n_2 y los números de aceptación c_1, c_2 y c_3 deben satisfacer los riesgos deseados del consumidor y del productor.

NOTAS Y COMENTARIOS

1. El uso de la distribución binomial en el muestreo de aceptación se basa en la suposición de que los lotes sean grandes. Si los lotes son pequeños, la distribución adecuada es la distribución hipergeométrica. Los expertos en el campo del control de calidad indican que la distribución de Poisson proporciona una buena aproximación en el muestreo de aceptación, cuando el tamaño de la muestra es por lo menos 16, el tamaño del lote es por lo menos 10 veces el tamaño de la muestra y p es menor a 0.10. Cuando se tienen muestras grandes se puede usar la aproximación normal a la distribución binomial.
2. En las tablas para muestreo MIL-ST-105D, a p_0 se le llama el nivel de calidad aceptable (AQL, por sus siglas en inglés) En algunas tablas de muestreo, a p_1 se le llama tolerancia de porcentaje de defectos en el lote (LTPD, por sus siglas en inglés) o el nivel de calidad rechazable (RQL. por sus siglas en inglés). En muchos de los planes de muestreo publicados también se usan índices de calidad como el nivel de calidad de indiferencia (IQL, por sus siglas en inglés) y el límite de calidad del promedio saliente (AOQL, por sus siglas en inglés). Textos más avanzados, listados en la bibliografía, proporcionan un análisis completo de estos índices.
3. En esta sección se presentó una introducción a los *planes de muestreo por atributos.* En estos planes cada artículo se clasifica como no defectuoso o defectuoso. En los *planes de muestreo de variables* se toma una muestra y se miden sus características de calidad. Por ejemplo, en el caso de joyería de oro, una medida de su calidad podría ser la cantidad de oro que contiene. Para decidir si aceptar o rechazar un lote podría emplearse un estadístico sencillo, como la cantidad promedio de oro en una muestra de joyas, el cual se compararía con la cantidad admitida de oro.

Ejercicios

Métodos



10. Determine la probabilidad de aceptar un lote que tiene 2% de defectos, si el plan de muestreo que se emplea es $n = 25$, $c = 0$. ¿Cuál será la probabilidad de aceptar el lote si el porcentaje de defectos es 6%?
11. Dado el plan de muestreo $n = 20$, $c = 0$, calcule el riesgo del productor en cada una de las situaciones siguientes.
    a. El porcentaje de defectos en el lote es 2%.
    b. El porcentaje de defectos en el lote es 6%.
12. Repita el ejercicio 11 con el plan de aceptación $n = 20$, $c = 1$. ¿Qué pasa con el riesgo del productor cuando aumenta el criterio de aceptación c? Explique.

Aplicaciones

13. Remítase al problema de KALI que se presentó en esta sección. El gerente de control de calidad requiere que el riesgo del productor sea 0.10 y $p_0 = 0.3$ y que el riesgo del consumidor sea 0.20 y $p_1 = 0.15$. En el plan de aceptación el tamaño de la muestra es 20 y el criterio de aceptación es 1. Resuelva las preguntas siguientes.
    a. ¿Cuál es el riesgo del productor, si el plan de aceptación es $n = 20$ y $c = 1$?
    b. ¿Cuál es el riesgo del consumidor, si el plan de aceptación es $n = 20$ y $c = 1$?
    c. ¿El plan de muestreo $n = 20$, $c = 1$ satisface los requisitos establecidos por el gerente de control de calidad? Analice.
14. Para inspeccionar un pedido de materia prima, recibido por una empresa, se piensa en usar muestras de tamaño 10, 15 y 20. Use las probabilidades binomiales de la tabla 5 del apéndice B para elegir un plan de muestreo con el riesgo del productor $\alpha = 0.03$ y $p_0 = 0.05$ y el riesgo del consumidor $\beta = 0.12$ y $p_1 = 0.30$.

15. Un fabricante de relojes le compra cristales de cuarzo a una empresa suiza. Estos cristales se surten en lotes de 1 000 piezas. Para el proceso de muestreo de aceptación se toman 20 cristales elegidos aleatoriamente.
    a. Trace las curvas características para los criterios de aceptación 0, 1 y 2.
    b. Si $p_0 = 0.01$ y $p_1 = 0.08$, ¿cuáles son los riesgos del consumidor y del productor con cada uno de los planes de muestreo del inciso a?

Resumen

En este capítulo se vio cómo usar los métodos estadísticos como ayuda en el control de calidad. Primero se presentaron las cartas de control $\bar{x}$, *R, p* y *np* que ayudan en el monitoreo de la calidad de los procesos. En cada una de estas cartas se establecen límites de control, se toman muestras de manera periódica y se grafican los puntos correspondientes en las cartas de control. Si hay puntos que caigan fuera de los límites de control, eso indica que el proceso está fuera de control y que se deben tomar las medidas correctivas correspondientes. También, algunos patrones que suelen seguir los datos, dentro de los límites de control, pueden indicar posibles problemas de control de calidad y sugerir que se tomen medidas correctivas.

Además se vio la técnica conocida como muestreo de aceptación. Esta técnica consiste en tomar una muestra e inspeccionarla. Un lote se acepta o rechaza, de acuerdo con el número de defectos encontrados en la muestra. El tamaño de la muestra y el criterio de aceptación pueden ajustarse para controlar tanto el riesgo del productor (error tipo I) como el riesgo del consumidor (error tipo II).

Glosario

Calidad total (CT) Sistema para mejorar la satisfacción del cliente y bajar los costos reales mediante una estrategia de mejoramiento y aprendizaje continuos.

Seis Sigma Metodología que emplea mediciones y análisis estadísticos para alcanzar un nivel de calidad tan bueno que en cada millón de operaciones no haya más de 3.4 defectos.

Control de calidad Una serie de inspecciones y mediciones que determinan si se han alcanzado los estándares de calidad establecidos.

Causas asignables Variaciones en los resultados de un proceso debidas a desgaste de la maquinaria, mala calidad de los materiales, errores de los operadores, etc. Cuando se detecta que la variación se debe a causas asignables, es necesario tomar medidas correctivas.

Causas comunes Variaciones naturales o normales en los resultados de un proceso, que son debidas, únicamente, a la casualidad. Cuando las variaciones en los resultados de un proceso se deben a causas comunes, no es necesario tomar ninguna medida.

Carta de control Herramienta gráfica que se usa como ayuda para determinar si un proceso está bajo control o fuera de control.

Carta $\bar{x}$ Carta de control que se usa cuando se mide la calidad de los resultados de un proceso en términos de longitud, peso, temperatura, etcétera.

Carta *R* Carta de control que se usa cuando se mide la calidad de los resultados de un proceso en términos del rango de una variable.

Carta *p* Carta de control que se usa cuando se mide la calidad de los resultados de un proceso en términos de la proporción de defectos.

Carta *np* Carta de control que se usa para monitorear la calidad de los resultados de un proceso mediante el número de artículos defectuosos.

Lote Conjunto de artículos, como pedidos recibidos de materias primas, de piezas o de bienes terminados para obtener un producto final.

Muestreo de aceptación Método estadístico en el que el número de elementos defectuosos que se encuentran en una muestra se usa para determinar si se acepta o se rechaza un lote.

Riesgo del productor Es el riesgo de rechazar un lote de buena calidad; error tipo I.

Riesgo del consumidor Es el riesgo de aceptar un lote de mala calidad; error tipo II.

Criterio de aceptación El número máximo de artículos defectuosos que se pueden encontrar en una muestra, para que a pesar de ello se acepte un lote.

Curva característica de operación (CO) Gráfica en la que se muestra la probabilidad de aceptar un lote, en función del porcentaje de artículos defectuosos encontrados en el lote. Esta curva se emplea para determinar si un plan de muestreo de aceptación satisface las exigencias tanto del consumidor como del productor.

Plan de muestreo múltiple Una forma de muestreo de aceptación en la que se usa más de una muestra o etapas. De acuerdo con el número de artículos defectuosos que se encuentran en una muestra, se decide si aceptar el lote, rechazar el lote o seguir con el muestreo.

Fórmulas clave

Error estándar de la media

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{(20.1)}$$

Límites de control en una carta $\bar{x}$: media del proceso y desviación estándar conocidas

$$\text{UCL} = \mu + 3\sigma_{\bar{x}} \qquad \textbf{(20.2)}$$
$$\text{LCL} = \mu - 3\sigma_{\bar{x}} \qquad \textbf{(20.3)}$$

Media muestral general

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k} \qquad \textbf{(20.4)}$$

Rango promedio

$$\bar{R} = \frac{R_1 + R_2 + \cdots + R_k}{k} \qquad \textbf{(20.5)}$$

Límites de control en una carta $\bar{x}$: media del proceso y desviación estándar desconocidas

$$\bar{\bar{x}} \pm A_2\bar{R} \qquad \textbf{(20.8)}$$

Límites de control en una carta R

$$\text{UCL} = \bar{R}D_4 \qquad \textbf{(20.14)}$$
$$\text{LCL} = \bar{R}D_3 \qquad \textbf{(20.15)}$$

Error estándar de la proporción

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad \textbf{(20.16)}$$

Límites de control en una carta p

$$\text{UCL} = p + 3\sigma_{\bar{p}} \qquad \textbf{(20.17)}$$
$$\text{LCL} = p - 3\sigma_{\bar{p}} \qquad \textbf{(20.18)}$$

Límites de control en una carta np

$$\text{UCL} = np + 3\sqrt{np(1-p)} \qquad \textbf{(20.19)}$$
$$\text{LCL} = np - 3\sqrt{np(1-p)} \qquad \textbf{(20.20)}$$

Distribución de probabilidad binomial para un muestreo de aceptación

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)} \tag{20.21}$$

Ejercicios complementarios

16. En un proceso de producción que se considera bajo control, las muestras de 5 elementos arrojaron las medias muestrales siguientes.

95.72	95.24	95.18
95.44	95.46	95.32
95.40	95.44	95.08
95.50	95.80	95.22
95.56	95.22	95.04
95.72	94.82	95.46
95.60	95.78	

a. Con base en estos datos dé la estimación de la media cuando el proceso está bajo control.
b. Si la desviación estándar del proceso es $\sigma = 0.50$, elabore la carta de control $\bar{x}$ de este proceso de producción. Como media del proceso considere la estimación obtenida en el inciso a.
c. ¿Alguna de las 20 medias muestrales indica que el proceso está fuera de control?

17. En un proceso los pesos de llenado tienen una distribución normal, la media es de 350 gramos y la desviación estándar de 15 gramos.
a. Obtenga los límites de control de la carta $\bar{x}$ para muestras de tamaño 10, 20 y 30.
b. ¿Qué pasa con los límites de control a medida que aumenta el tamaño de la muestra?
c. ¿Qué pasa cuando se comete un error tipo I?
d. ¿Qué pasa cuando se comete un error tipo II?
e. Calcule la probabilidad de cometer un error tipo I con muestras de los tamaños: 10, 20 y 30.
f. En las cartas de control, ¿qué ventaja tiene incrementar los tamaños de la muestra? ¿Qué probabilidad de error se reduce a medida que se incrementa el tamaño de la muestra?

18. A partir de muestras de tamaño 5 se obtuvo $\bar{\bar{x}} = 5.42$ y $\bar{R} = 2.0$. Calcule los límites de control de las cartas $\bar{x}$ y R, y estime la desviación estándar del proceso.

19. Los siguientes datos de control de calidad se obtuvieron en un proceso de fabricación de la empresa Kensport Chemical. Estos datos son temperaturas en grados centígrados medidas en cinco momentos del ciclo de producción. La empresa está interesada en emplear cartas de control para monitorear la temperatura durante su proceso de fabricación. Construya las cartas $\bar{x}$ y R . ¿Qué se puede concluir acerca de la calidad del proceso?

Muestra	$\bar{x}$	***R***	**Muestra**	$\bar{x}$	***R***
1	95.72	1.0	11	95.80	0.6
2	95.24	0.9	12	95.22	0.2
3	95.18	0.8	13	95.56	1.3
4	95.44	0.4	14	95.22	0.5
5	95.46	0.5	15	95.04	0.8
6	95.32	1.1	16	95.72	1.1
7	95.40	0.9	17	94.82	0.6
8	95.44	0.3	18	95.46	0.5
9	95.08	0.2	19	95.60	0.4
10	95.50	0.6	20	95.74	0.6

20. Los siguientes datos se obtuvieron del proceso de producción de Master Blend Coffee y son los pesos de llenado de latas de café de 3 libras. Con estos datos construya las tablas $\bar{x}$ y R. ¿Qué se puede concluir acerca de la calidad de este proceso de producción?



	Observaciones				
Muestra	**1**	**2**	**3**	**4**	**5**
1	3.05	3.08	3.07	3.11	3.11
2	3.13	3.07	3.05	3.10	3.10
3	3.06	3.04	3.12	3.11	3.10
4	3.09	3.08	3.09	3.09	3.07
5	3.10	3.06	3.06	3.07	3.08
6	3.08	3.10	3.13	3.03	3.06
7	3.06	3.06	3.08	3.10	3.08
8	3.11	3.08	3.07	3.07	3.07
9	3.09	3.09	3.08	3.07	3.09
10	3.06	3.11	3.07	3.09	3.07

21. Considere la situación siguiente y diga si hay razones para preocuparse por la calidad del proceso.
   a. En una carta p, se tiene LCL = 0 y UCL = 0.068. Cuando el proceso está bajo control, la proporción de defectos es 0.033. En esta carta de control grafique los siguientes siete resultados muestrales: 0.035, 0.062, 0.055, 0.049, 0.058, 0.066 y 0.055. Analice.
   b. En una carta $\bar{x}$, se tiene LCL = 22.2 y UCL = 24.5. Cuando el proceso está bajo control, la media es $\mu = 23.35$. En esta carta grafique los siguientes siete resultados muestrales: 22.4, 22.6, 22.65, 23.2. 23.4, 23.85 y 24.1. Analice.

22. A veces al mes, los gerentes de 1 200 comercios minoristas hacen pedidos de reabastecimiento a la casa matriz. De acuerdo con la experiencia se sabe que cerca de 4% de los pedidos que se surten presentan algún error, como envío de una mercancía distinta a la solicitada, envío de una cantidad distinta a la solicitada o, simplemente, no se surte la mercancía solicitada. Cada mes se toman muestras de 200 hojas de pedido y se verifica si fueron surtidas con precisión.
   a. Construya la carta de control correspondiente a esta situación.
   b. En los datos de los últimos seis meses las cantidades de pedidos con uno o más errores son: 10, 15, 6, 13, 8 y 17. Grafique estos datos en la carta de control. ¿Qué dice esta carta acerca de este proceso?

23. Se tiene el plan de muestreo de aceptación $n = 10$, $c = 2$. Suponga que $p_0 = 0.05$ y $p_1 = 0.20$.
   a. Para este plan de muestreo de aceptación calcule el riesgo del consumidor.
   b. ¿Estará inconforme el productor, el consumidor o ambos con el plan de muestreo propuesto?
   c. ¿Recomendaría usted alguna modificación al plan de muestreo? ¿Cuál?

24. Se elaboró el plan de muestreo de aceptación $n = 15$, $c = 1$ con un riesgo para el productor de 0.075.
   a. ¿Cuál es al valor de $p_{0,}$ 0.01, 0.02, 0.03, 0.04 o 0.05? ¿Qué significa este valor?
   b. ¿Cuál es el riesgo del consumidor con este plan si $p_1 = 0.25$?

25. Una empresa produce carne enlatada. Sea p la proporción de lotes que no satisfacen las especificaciones de calidad del producto. Como plan de muestreo de aceptación se va a usar $n = 25$, $c = 0$.
   a. Calcule los puntos que en la curva característica de operación corresponden a $p = 0.01$, 0.03, 0.10 y 0.20.
   b. Grafique la curva característica de operación.
   c. ¿Cuál es la probabilidad de que con este plan de muestreo de aceptación se rechace un lote que contenga 0.01 defectos?

Apéndice 20.1 Cartas de control con Minitab


Jensen

En este apéndice se describen los pasos necesarios para generar cartas de control con Minitab, a partir de los datos muestrales de la empresa Jensen, presentados en la tabla 20.1. Los números de las muestras se encuentran en la columna C1, la primera observación se encuentra en la columna C2, la segunda, en la columna C3 y así sucesivamente. Mediante los pasos siguientes, Minitab genera simultáneamente las cartas $\bar{x}$ y R.

Paso 1. Seleccionar el menú **Stat**
Paso 2. Elegir **Control Charts**
Paso 3. Elegir **Variables Charts for Subgroups**
Paso 4. Elegir **Xbar-R**
Paso 5. Cuando aparezca el cuadro de diálogo Xbar-R Chart:
Seleccionar **Observations for a subgroup are in one row of columns**
En el cuadro que se encuentra debajo ingresar C2-C6
Seleccionar **Xbar-R Options**
Paso 6. Cuando aparezca el cuadro de diálogo Xbar-R Options:
Seleccionar la pestaña **Tests**
Elegir **One point > 3.0 standard deviations from center line***
Clic en **OK**
Paso 7. Cuando aparezca el cuadro de diálogo Xbar-R Chart:
Clic en **OK**

En los resultados de Minitab aparecerán, juntas, la carta R y la carta $\bar{x}$. En el paso 3 de este procedimiento aparecen diversas opciones que permiten el acceso a diferentes tipos de cartas de control. Por ejemplo, se puede seleccionar que las cartas $\bar{x}$ y R aparezcan por separado. Otras de las opciones son obtener una carta *p*, una carta *np*, etcétera.

* Minitab proporciona otras pruebas para detectar causas especiales de variación y situaciones fuera de control. El usuario puede elegir simultáneamente varias de estas pruebas

CAPÍTULO 21



Análisis de decisión

CONTENIDO



LA ESTADÍSTICA *en* LA PRÁCTICA

OHIO EDISON COMPANY*
AKRON, OHIO

Ohio Edison Company es una empresa de FirstEnergy Corporation. Ohio Edison y su subsidiaria Pensilvania Power Company, suministran energía eléctrica a más de 1 millón de usuarios en el centro y noreste de Ohio y en el oeste de Pensilvania. La mayor parte de la electricidad la generan mediante plantas de combustión de carbón. Debido a los requerimientos de control de la contaminación, Ohio Edison se embarcó en un programa para renovar su equipo para el control de la contaminación en la mayor parte de sus plantas generadoras.

Para satisfacer los nuevos límites de emisión de dióxido de azufre en una de sus plantas más grandes, Ohio Edison decidió quemar carbono con bajo contenido de azufre en cuatro de las unidades más pequeñas de la planta e instalar filtros de tela en esas unidades para controlar la emisión de partículas. Los filtros de tela usan miles de bolsas de tela para retener las partículas y funcionan de manera muy parecida a las aspiradoras caseras.

En las tres unidades más grandes de la planta se consideró la posibilidad de quemar carbón de medio a alto contenido de azufre. Estudios preliminares redujeron las opciones de equipo para retención de partículas para estas unidades más grandes a filtros de tela y precipitadores electrostáticos (que eliminan las partículas que se encuentran suspendidas en el humo al pasarlo a través de un fuerte campo eléctrico). Entre las incertidumbres al tomar una decisión final estaba la manera en que pueden interpretarse algunas leyes y normas, los potenciales cambios en las leyes y normas sobre calidad del aire y las fluctuaciones en los costos de construcción.

Debido a la complejidad del problema, el alto grado de incertidumbre relacionado con los factores que afectaban la decisión y el impacto de los costos para Ohio Edison, se empleó el análisis de decisión. Entonces se elaboró una descripción gráfica del problema, a la que se conoce como árbol de decisión. Para evaluar los resultados mostrados por el árbol de decisión se consideraron las necesidades de ingreso anual de las tres unidades grandes por el resto de su vida útil. Las necesidades de ingreso anual eran los dineros que debían obtenerse de los usuarios para recobrar el costo de la instalación del nuevo equipo para el control de la contaminación. Mediante el análisis del árbol de decisión se llegó a las conclusiones siguientes.



Las plantas de Ohio Edison suministran energía eléctrica a más de 1 millón de usuarios. © Getty Images/PhotoDisc.

- El valor esperado de las necesidades de ingreso anual del precipitador electrostático era aproximadamente 1 millón de dólares inferior al de las necesidades para los filtros de tela.
- Los filtros de tela tenían una probabilidad mayor de necesidades de ingreso alto que los precipitadores electrostáticos.
- Los precipitadores electrostáticos tenían una probabilidad de casi 0.8 de tener necesidades de ingreso anual menores.

Estas conclusiones llevaron a Ohio Edison a decidirse por los precipitadores electrostáticos para las unidades generadoras en cuestión. Si no se hubiera empleado el análisis de decisión, la toma de decisión se hubiera basado principalmente en el costo de la inversión, lo cual hubiera llevado a resolverse por el equipo de los filtros de tela. El uso del análisis de decisión permitió identificar la opción que tenía tanto las menores necesidades de ingreso esperadas como el menor riesgo.

En este capítulo se verá la metodología del análisis de decisión empleada por Ohio Edison. La atención se centra en mostrar cómo el análisis de decisión permite identificar la mejor alternativa ante un panorama lleno de riesgos por eventos futuros.

* Los autores agradecen a Thomas J. Madden y a M. S. Hyrnick de Ohio Edison por proporcionar este artículo para *La estadística en la práctica.*

El análisis de decisión se usa para elaborar una estrategia óptima de decisión ante diversas alternativas y ante un conjunto de eventos futuros inciertos y llenos de riesgos. Para iniciar el estudio del análisis de decisión se emplearán problemas de decisión en los que las alternativas de decisión y los inciertos eventos futuros sean razonablemente pocos. Se presentarán las tablas de recompensa con objeto de dar una estructura a los problemas de decisión. Después se introducirán los árboles de decisión con objeto de mostrar la naturaleza secuencial de los problemas. Los árboles de decisión se usan para analizar problemas más complejos y para identificar una secuencia óptima de decisiones, a la que se le conoce como estrategia óptima de decisión. En la última sección se muestra cómo usar el teorema de Bayes, presentado en el capítulo 4, para calcular las probabilidades de las ramas de los árboles de decisión. En el sitio de la red de ASW, http://asw.swlearning.com., se proporciona TreePlan, el complemento de Excel para el análisis de decisiones y se dan las indicaciones para su empleo.

En el apéndice 21.1 se presenta un ejemplo del software para análisis de decisión TreePlan.



21.1 Formulación del problema

El primer paso en el análisis de decisión es formular el problema. Se empieza por hacer un planteamiento verbal del mismo. Después se identifican las alternativas de decisión, los eventos futuros inciertos, conocidos como **eventos aleatorios**, y las **consecuencias** de cada combinación de una alternativa de decisión con uno de los resultados del evento aleatorio. Como ejemplo se considerará un proyecto de construcción de la empresa Pittsburgh Development Corporation.

Pittsburgh Development Corporation (PDC) compró un terreno para construir un lujoso complejo de condominios. El lugar tiene una vista espectacular sobre el centro de Pittsburgh y el Golden Triangle formado por los ríos Allegheny y Monongahela que se unen para formar el río Ohio. PDC desea vender cada condominio en un precio entre $300 000 y $1 400 000.

Para empezar, PDC ha encargado tres proyectos arquitectónicos de distintos tamaños: uno de 30 condominios, otro de 60 condominios y el tercero de 90 condominios. El éxito del proyecto dependerá tanto del tamaño del complejo como del evento aleatorio de la demanda que pueda haber por los condominios. El problema de decisión de PDC es elegir el tamaño del complejo que conduzca a las mayores ganancias, dada la incertidumbre relativa en la demanda de los condominios.

De acuerdo con este planteamiento del problema, es claro que la decisión es el tamaño adecuado del condominio. PDC tiene las siguientes tres alternativas para su decisión:

d_1 = un complejo pequeño de 30 condominios
d_2 = un complejo mediano de 60 condominios
d_3 = un complejo grande de 90 condominios

Un factor importante en la elección de la mejor alternativa es la incertidumbre relacionada con el evento aleatorio de la demanda que pueda haber por el condominio. Al preguntarse por las posibilidades de demanda, el presidente de PDC reconoce que existe una amplia gama, pero considera la ocurrencia de dos eventos aleatorios: una demanda alta y una demanda baja.

En el análisis de decisión, a los posibles resultados de un evento aleatorio se les conoce como **estados**. Los estados se definen de tal manera que uno y sólo uno de los estados pueda presentarse. En el problema de PDC, el evento aleatorio de la demanda de los condominios tiene dos estados:

s_1 = una demanda alta de los condominios
s_2 = una demanda baja de los condominios

Los directivos elegirán, primero, una alternativa de decisión (tamaño del complejo), después seguirá un estado (demanda por los condominios) y por último se tendrá una consecuencia. En este caso, la consecuencia son las ganancias que obtendrá PDC.

Tablas de recompensa

Dadas las tres alternativas de decisión y los dos estados, ¿qué tamaño de condominio debe elegirse? Para responder esta pregunta, PDC necesita conocer las consecuencias de cada una de las combinaciones de alternativa de decisión y un estado. En el análisis de decisión, a cada una de las consecuencias de la combinación de una alternativa de decisión y un estado se le conoce como **recompensa**. A la tabla en la que se muestran las recompensas de todas las combinaciones de alternativa de decisión y un estado, se le conoce como **tabla de recompensas**.

Las recompensas se pueden expresar en términos de ganancias, costos, tiempo, distancia o cualquier otra cantidad apropiada para el problema de decisión que se analice.

Como PDC desea elegir el tamaño de complejo que le proporcione mayores ganancias, las ganancias se usarán como consecuencia. En la tabla 21.1 se muestra la tabla de recompensa, que expresa las ganancias en millones de dólares. Observe que si el tamaño del condominio es mediano y la demanda es alta las ganancias serán de \$14 millones. La recompensa correspondiente a cada combinación de una alternativa de decisión i y un estado j se denotará V_{ij}. Así, de acuerdo con la tabla 21.1, $V_{31} = 20$ significa que habrá una recompensa de \$20 millones si la decisión es construir un complejo grande (d_3) y la demanda que se presenta es alta (s_1). De manera similar $V_{32} = -9$ significa que habrá una pérdida de \$9 millones si la decisión es construir un complejo grande (d_3) y la demanda que se presenta baja (s_2).

Árboles de decisión

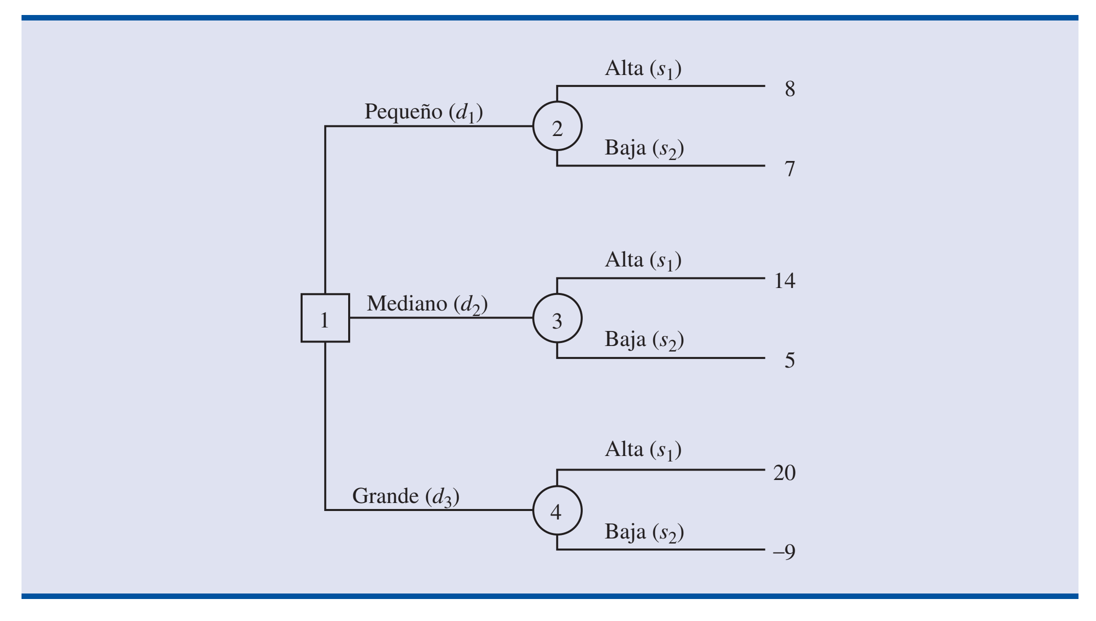
En un **árbol de decisión** se muestra gráficamente el carácter secuencial del proceso de toma de decisión. En la figura 21.1 se presenta el árbol de decisión para el caso del problema de PDC; en el árbol de decisión se muestra la progresión lógica o natural en el tiempo. Primero, PDC tendrá que tomar una decisión respecto al tamaño del complejo de condominios (d_1, d_2 o d_3). Después de llevar a cabo lo que se haya decidido, se dará el estado s_1 o el estado s_2. El número que aparece en cada punto terminal del árbol es la recompensa correspondiente a la secuencia dada. Por ejemplo, la recompensa 8, que es la que se encuentra más arriba, significa que se espera una ganancia de \$8 millones si PDC construye un complejo pequeño (d_1) y la demanda resulta ser alta (s_1). La recompensa siguiente, que es 7, significa que se espera una ganancia de \$7 millones si PDC construye un complejo pequeño (d_1) y la demanda resulta ser baja (s_2). De esta manera, en este árbol de decisión se muestran gráficamente las secuencias de alternativas de decisión y estados con los que se llega a las seis recompensas posibles.

El árbol de decisión de la figura 21.1 tiene cuatro **nodos**, numerados del 1 al 4, que representan las decisiones y los eventos aleatorios. Para representar **nodos de decisión** se emplean cuadrados y para representar **nodos aleatorios** se emplean círculos. Así, el nodo 1 es un nodo de decisión, y los nodos 2, 3 y 4 son nodos aleatorios. Las **ramas** que salen del nodo de decisión son las alternativas de decisión. Las ramas que salen de cada nodo aleatorio son estados. Las recompensas aparecen al final de las de los estados. Ahora se vuelve a la pregunta: ¿cómo puede,

TABLA 21.1 TABLA DE RECOMPENSA PARA EL PROYECTO DEL CONDOMINIO DE PDC (RECOMPENSAS EN MILLONES DE DÓLARES)

	Estado	
Alternativa de decisión	Demanda alta s_1	Demanda baja s_2
Complejo pequeño, d_1	8	7
Complejo mediano, d_2	14	5
Complejo grande, d_3	20	−9

FIGURA 21.1 ÁRBOL DE DECISIÓN PARA EL PROYECTO DEL CONDOMINIO DE PDC (RECOMPENSAS EN MILLONES DE DÓLARES)



la persona que toma la decisión, usar la información de la tabla de recompensa o del árbol de decisión para elegir la mejor alternativa de decisión?

NOTAS Y COMENTARIOS

1. Los expertos en solución de problemas coinciden en que el primer paso en la resolución de un problema complejo es descomponerlo en una serie de subproblemas menores. Los árboles de decisión sirven para mostrar cómo descomponer el problema y también para mostrar el carácter secuencial del proceso de decisión.
2. Las personas suelen ver un mismo problema desde diferentes perspectivas. Por tanto, la discusión sobre la elaboración de un árbol de decisión puede proporcionar mayor claridad acerca del problema.



21.2 Toma de decisiones con probabilidades

Una vez definidas las alternativas de decisión y los estados de los eventos aleatorios, se determinan las probabilidades de los estados. Para determinar estas probabilidades se puede usar cualquiera de los métodos estudiados en el capítulo 4, el método clásico, el método de las frecuencias relativas o el método subjetivo. A continuación se muestra cómo usar, una vez determinadas las probabilidades, el **método del valor esperado** para identificar la mejor alternativa de decisión o la decisión recomendada para el problema dado.

Método del valor esperado

Se empezará por definir el valor esperado de una alternativa de decisión. Sea

$$N = \text{cantidad de estados}$$
$$P(s_j) = \text{probabilidad del estado } s_j$$

Como únicamente puede presentarse uno y sólo uno de los N estados, las probabilidades deben satisfacer dos condiciones:

Las probabilidades asignadas a los estados deben satisfacer los requisitos básicos de la asignación de probabilidades presentados en el capítulo 4.

$$P(s_j) \geq 0 \qquad \text{para todos los estados} \tag{21.1}$$

$$\sum_{j=1}^{N} P(s_j) = P(s_1) + P(s_2) + \cdots + P(s_N) = 1 \tag{21.2}$$

El **valor esperado (VE)** de una alternativa de decisión d_i es el siguiente.

VALOR ESPERADO

$$\text{VE}(d_i) = \sum_{j=1}^{N} P(s_j)V_{ij} \tag{21.3}$$

donde

V_{ij} = valor de la recompensa para la alternativa de decisión d_i y el estado s_j.

Es decir, el valor esperado de una alternativa de decisión es la suma de las recompensas ponderadas que hay para esa alternativa de decisión. El peso de ponderación para una recompensa es la probabilidad de que dicha recompensa ocurra. Para ver cómo se emplea el método del valor esperado se vuelve al problema de PDC.

PDC ve con optimismo el potencial del lujoso complejo de condominios. Este optimismo lo lleva a una evaluación inicial, mediante el método de probabilidad subjetiva, y asigna la probabilidad 0.8 al evento de que la demanda sea alta (s_1) y 0.2 al evento de que la demanda sea baja (s_2). Por tanto, $P(s_1) = 0.8$ y $P(s_2) = 0.2$. Con los valores de recompensa de la tabla 21.1 y la ecuación (21.3), el valor esperado de cada una de las tres alternativas de decisión se calcula como sigue.

$$\begin{aligned}
\text{VE}(d_1) &= 0.8(8) + 0.2(7) = 7.8 \\
\text{VE}(d_2) &= 0.8(14) + 0.2(5) = 12.2 \\
\text{VE}(d_3) &= 0.8(20) + 0.2(-9) = 14.2
\end{aligned}$$

De esta manera, con el método del valor esperado, se encuentra que el complejo grande, cuyo valor esperado es 14.2 millones de dólares, es la decisión recomendada.

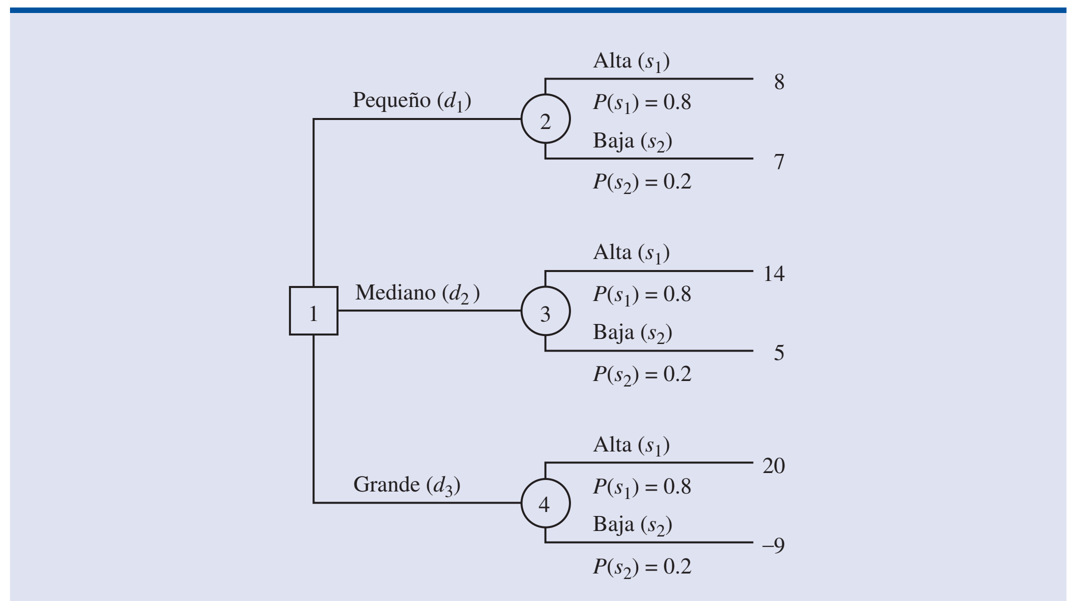
Los cálculos para identificar la alternativa de decisión que tiene el mejor valor esperado pueden realizarse en un árbol de decisión. En la figura 21.2 se muestra el árbol de decisión del problema de PDC con las probabilidades en las ramas de los estados. Al recorrer el árbol de decisión de atrás para adelante, se calcula primero el valor esperado en cada nodo aleatorio; es decir, para cada nodo aleatorio se ponderan las posibles recompensas al multiplicarlas por probabilidad de ocurrencia. De esta manera se obtiene el valor esperado para los nodos 2, 3 y 4, como se muestra en la figura 21.3.

Como el que toma la decisión controla la rama que sale del nodo 1 de decisión y como se trata de maximizar las ganancias esperadas, la mejor alternativa de decisión en el nodo 1 es d_3. Por tanto, el análisis del árbol de decisión lleva a la recomendación de d_3, cuyo valor esperado es $14.2 millones. Observe que con el método del valor esperado en conjunción con la tabla de recompensas se obtiene la misma recomendación.

Para la construcción de árboles de decisión más complejos existen paquetes de software.

Los problemas de decisión pueden ser bastante más complejos que el problema de decisión de PDC, pero siempre que la cantidad de alternativas de decisión y de estados sea razonable, se podrá emplear el método de árbol de decisión aquí descrito. Primero se dibuja un árbol de decisión que consista únicamente en los nodos de decisión, los nodos aleatorios y las ramas que describen el carácter secuencial del problema. Si se usa el método del valor esperado, el paso siguiente es determinar las probabilidades de cada uno de los estados y calcular el valor esperado

FIGURA 21.2 ÁRBOL DE DECISIÓN PARA EL PROBLEMA DE PDC CON LAS PROBABILIDADES EN LAS RAMAS DE ESTADO

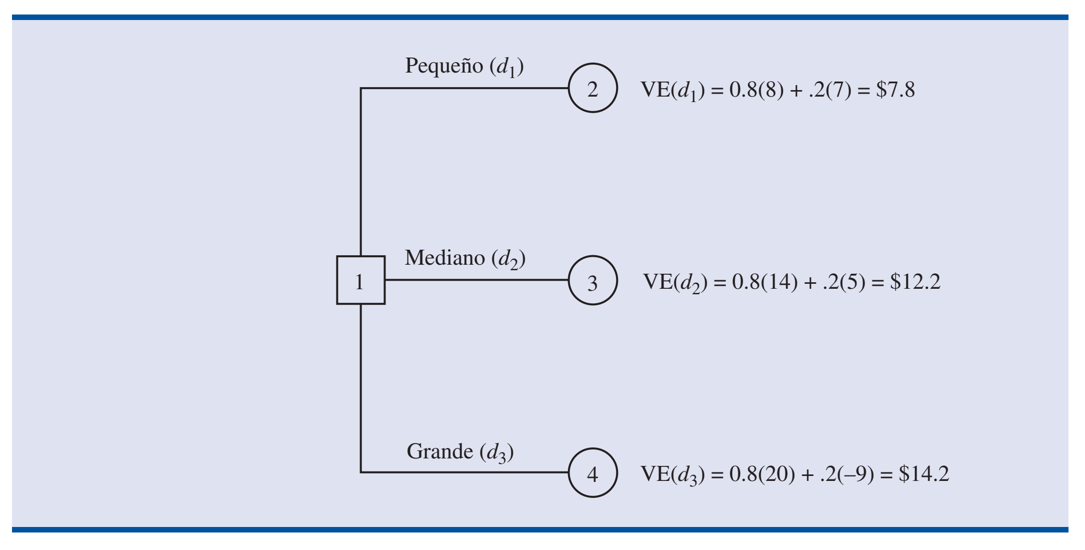


en cada nodo aleatorio. A continuación se elige la rama de decisión que lleva al nodo aleatorio que tenga el mayor valor esperado. La alternativa de decisión correspondiente a esta rama es la decisión recomendada.

Valor esperado de la información perfecta

Suponga que PDC tiene la oportunidad de realizar un estudio de investigación de mercado que le ayudará a evaluar el interés del público por el proyecto del condominio y que proporcionará a los directivos información para mejorar su evaluación de las probabilidades de los estados. Para determinar el valor potencial de esta información, se comenzará por suponer que el estudio puede proporcionar *información perfecta* sobre los estados; es decir, se acepta, por el momento, que

FIGURA 21.3 APLICACIÓN DEL MÉTODO DEL VALOR ESPERADO MEDIANTE UN ÁRBOL DE DECISIÓN



PDC podría determinar con certeza, antes de tomar una decisión, qué estado va a ocurrir. Para aprovechar esta información perfecta, se elaborará una estrategia de información, la cual seguirá PDC una vez que sepa qué estado se presenta. Una estrategia de decisión es simplemente una regla de decisión que especifica la alternativa de decisión a elegir una vez que se cuente con más información.

Como ayuda para determinar la estrategia de decisión de PDC, en la tabla 21.2 se reproduce la tabla de recompensas de PDC. Observe que si PDC supiera con certeza que el estado s_1 fuera a ocurrir, la mejor alternativa de decisión sería d_3, cuya recompensa es de 20 millones de dólares. De manera similar, si PDC supiera con certeza que el estado s_2 es el que va a ocurrir, la mejor alternativa de decisión sería d_1, cuya recompensa es 7 millones de dólares. Por tanto, se puede establecer la estrategia óptima de decisión si PDC llega a contar con la información perfecta como sigue:

Si s_1, se elige d_3 y se obtiene una recompensa de $20 millones.

Si s_2, se elige d_1 y obtiene una recompensa de $7 millones.

¿Cuál es el valor esperado con esta estrategia de decisión? Para calcular el valor esperado con información perfecta se vuelve a las probabilidades originales de los estados: $P(s_1) = 0.8$ y $P(s_2) = 0.2$. Por tanto, hay una probabilidad de 0.8 de que la información perfecta indique el estado s_1 y la decisión alternativa resultante d_3 proporcionará $20 millones de ganancias. De manera similar, si 0.2 es la probabilidad del estado s_2, la alternativa óptima de decisión d_1 proporcionará $7 millones de ganancia. De esta manera, con la ecuación (21.3), el valor esperado de la estrategia de decisión basada en la información perfecta es

$$0.8(20) + 0.2(7) = 17.4$$

A este valor esperado de $17.4 millones se le conoce como el *valor esperado con información perfecta* (VEcIP).

Ya antes, en esta sección, se había indicado que la decisión recomendada usando el método del valor esperado era la alternativa de decisión d_3, con un valor esperado de $14.2 millones. Como la recomendación de esta decisión y el cálculo del valor esperado se hizo sin la ventaja de la información perfecta, a los $14.2 millones de dólares se les conoce como *valor esperado sin información perfecta* (VEsIP).

El valor de que PDC conozca el nivel de aceptación en el mercado antes de elegir una alternativa de decisión es $3.2 millones.

El valor esperado con información perfecta es $17.4 millones y el valor esperado sin información perfecta es $14.2 millones; por tanto, el valor esperado de la información perfecta (VEIP) es $17.4 − $14.2 = $3.2 millones. En otras palabras, $3.2 millones representan el valor esperado adicional que se puede obtener si se cuenta con información perfecta acerca del estado. En general, un estudio de investigación de mercado no proporciona información "perfecta"; pero si el estudio de mercado es bueno, la información obtenida bien puede valer una buena porción de los $3.2 millones. Dado que el VEIP es de $3.2 millones, PDC puede considerar seriamente un estudio de investigación de mercado con objeto de obtener más información acerca del estado.

TABLA 21.2 TABLA DE RECOMPENSA PARA EL PROYECTO DE LOS CONDOMINIOS DE PDC (MILLONES DE DÓLARES)

	Estado	
Alternativa de decisión	**Demanda alta s_1**	**Demanda baja s_2**
Complejo pequeño, d_1	8	7
Complejo mediano, d_2	14	5
Complejo grande, d_3	20	−9

El **valor esperado de la información perfecta** se calcula en general como sigue:

VALOR ESPERADO DE LA INFORMACIÓN PERFECTA

$$\text{VEIP} = |\text{VEcIP} - \text{VEsIP}| \tag{21.4}$$

donde

VEIP = valor esperado de la información perfecta
VEcIP = valor esperado con información perfecta acerca del estado
VEsIP = valor esperado sin información perfecta acerca del estado

Observe el papel del valor absoluto en la ecuación (21.4). En problemas de minimización, la información ayuda a reducir y bajar los costos; de manera que el valor esperado con información perfecta es menor o igual al valor esperado sin información perfecta. En este caso, VEIP es la magnitud de la diferencia entre VEcIP y VEsIP o el valor absoluto de la diferencia como se muestra en la ecuación (21.4).

Ejercicios

Métodos



1. En la tabla de recompensa siguiente se muestran las ganancias en un problema de análisis de decisión en el que se tienen dos alternativas de decisión y tres estados.

Alternativas de decisión	Estados s_1	s_2	s_3
d_1	250	100	25
d_2	100	100	75

   a. Construya un árbol de decisión para este problema.
   b. Suponga que la persona que debe tomar la decisión obtiene las probabilidades $P(s_1) = 0.65$, $P(s_2) = 0.15$ y $P(s_3) = 0.20$. Emplee el método del valor esperado para determinar la mejor decisión.

2. Una persona que debe tomar una decisión y que se encuentra ante cuatro alternativas de decisión y cuatro estados elabora la tabla de recompensa siguiente:

Alternativas de decisión	Estados s_1	s_2	s_3	s_4
d_1	14	9	10	5
d_2	11	10	8	7
d_3	9	10	10	11
d_4	8	10	11	13

   Esta persona obtiene información que le permite hacer las siguientes evaluaciones de las probabilidades: $P(s_1) = 0.5$, $P(s_2) = 0.2$, $P(s_3) = 0.2$ y $P(s_4) = 0.1$.
   a. Emplee el método del valor esperado para determinar la solución óptima.
   b. Suponga que las entradas en la tabla de recompensa son costos. Use el método del valor esperado para determinar la decisión óptima.

Aplicaciones



3. Hudson Corporation está en consideración de tres opciones para el procesamiento de sus datos: continuar con su propio personal, contratar una empresa externa para que lo haga (lo que se conoce como *outsourcing*) o una combinación de ambas cosas. El costo depende de la demanda futura. El costo anual de cada opción (en miles de dólares) depende de la demanda futura.

	Demanda		
Opciones	**Alta**	**Media**	**Baja**
Personal propio	650	650	600
Empresa externa	900	600	300
Combinación	800	650	500

a. Si las probabilidades para la demanda son 0.2, 0.5 y 0.3, respectivamente, ¿qué alternativa de decisión minimizará el costo del procesamiento de datos? ¿Cuál es el costo anual esperado de su recomendación?
b. ¿Cuál es el valor esperado de la información perfecta?

4. Myrtle Air Express ha decidido ofrecer un servicio directo de Cleveland a Myrtle Beach. Los directivos deben decidir entre un servicio de primera a precios normales usando la nueva flota de jet de la empresa o servicio de bajo precio usando los aviones regionales de menor capacidad. Es claro que la mejor elección depende de la reacción del mercado al servicio que ofrece Myrtle Air. Los administradores han elaborado estimaciones de la contribución a las ganancias que tendría cada tipo de servicio con base en dos niveles de demanda del servicio a Myrtle Beach: alta o baja. En la tabla siguiente se muestran las ganancias trimestrales estimadas (en miles de dólares).

	Demanda del servicio	
Servicio	**Alta**	**Baja**
De primera	$960	−$490
De bajo precio	$670	$320

a. ¿Qué es lo que hay que decidir? ¿Cuál es el evento aleatorio y cuál es la consecuencia? ¿Cuántas alternativas de decisión hay? ¿Cuántos resultados tiene el evento aleatorio?
b. Suponga que el gerente de Myrtle Air cree que la probabilidad de que la demanda sea alta es 0.7 y que la probabilidad de que la demanda sea baja es 0.3. Emplee el método del valor esperado para determinar cuál es la decisión óptima.
c. Suponga que la probabilidad de que la demanda sea alta es 0.8 y la probabilidad de que la demanda sea baja es 0.2. Emplee el método del valor esperado para determinar cuál es la decisión óptima.

5. La distancia de Potsdam a los grandes mercados y el limitado servicio aéreo ha impedido que este pueblo tenga un atractivo desarrollo industrial. Air Express, una importante empresa de servicio nocturno de transporte de paquetería, está considerando establecer un centro de distribución regional en Potsdam. Pero Air Express sólo establecerá este centro si el aeropuerto local aumenta la longitud de su pista. Otra empresa que pretende establecerse en esa localidad es Diagnostic Research, Inc. (DRI), uno de los principales productores de equipo para pruebas médicas. DRI pretende instalar una nueva planta de fabricación en el lugar. Para DRI no es condición el que se aumente la longitud de la pista aérea, pero la comisión de planificación local considera que eso serviría para convencer a DRI de abrir su nueva fábrica en Potsdam.

Si esta localidad aumenta la longitud de su pista aérea, la comisión de planificación local considera que se tendrían las probabilidades que se muestran en la tabla siguiente.

	Planta de DRI	Sin planta de DRI
Centro de Air Express	0.30	0.10
Sin centro de Air Express	0.40	0.20

Por ejemplo, la probabilidad de que Air Express establezca un centro de distribución y DRI construya una nueva planta en el lugar es 0.30.

Los ingresos anuales estimados para la localidad, una vez deducidos los costos de aumentar la longitud de la pista aérea, son los siguientes:

	Planta de DRI	Sin planta de DRI
Centro de Air Express	\$600 000	\$150 000
Sin centro de Air Express	\$250 000	−\$200 000

Si no se realiza el proyecto de expansión de la pista aérea, la comisión de planificación estima que la probabilidad de que DRI no establezca su nueva planta en ese lugar es 0.6, en este caso el ingreso anual estimado del lugar será de \$450 000. Si no se realiza el proyecto de expansión de la pista aérea y DRI no establece su nueva planta en ese lugar, el ingreso anual será \$0, ya que no se incurrirá en ningún gasto y no habrá ningún ingreso.

a. ¿Cuál es la decisión a tomar, cuál es el evento aleatorio y cuál es la consecuencia?
b. Calcule el ingreso anual esperado correspondiente a la alternativa de decisión de aumentar la longitud de la pista aérea.
c. Calcule el ingreso anual esperado correspondiente a la alternativa de decisión de no aumentar la longitud de la pista aérea.
d. ¿El pueblo debe elegir aumentar la longitud de la pista aérea? Explique.
e. Suponga que las probabilidades relacionadas con el aumento de la longitud de la pista aérea fueran las siguientes:

	Planta de DRI	Sin planta de DRI
Centro de Air Express	0.40	0.10
Sin centro de Air Express	0.30	0.20

¿Este cambio de las probabilidades tendría algún efecto sobre la decisión recomendada?

6. La empresa vitivinícola Seneca Hill Winery acaba de adquirir una propiedad con objeto de crear un nuevo viñedo. La dirección está considerando dos variedades de uva blanca: Chardonnay y Riesling. Con las uvas Chardonnay se produciría un vino Chardonnay seco y con las uvas Riesling se produciría un vino Riesling semiseco. Se necesitan aproximadamente cuatro años desde que se planta la uva hasta que puede ser cosechada. Este tiempo hace que se tenga una gran incertidumbre respecto a la demanda futura, y dificulta la decisión de qué tipo de uva sembrar. Se consideran tres posibilidades: sembrar únicamente uva Chardonnay, sembrar únicamente uva Riesling o sembrar ambas, Chardonnay y Riesling. Los directivos de Seneca han decidido que para los propósitos de la planeación bastará considerar únicamente dos posibilidades de deman-

da para cada tipo de vino: alta y baja. Al tener únicamente dos posibilidades para cada tipo de vino, es necesario evaluar cuatro probabilidades. Con ayuda de algunos pronósticos de publicaciones industriales, la dirección ha estimado las probabilidades siguientes.

	Demanda de Riesling	
Demanda de Chardonnay	**Baja**	**Alta**
Baja	0.05	0.50
Alta	0.25	0.20

Las proyecciones de ingresos muestran una contribución anual de $20 000 a las ganancias, si Seneca planta únicamente uva Chardonnay y la demanda de vino Chardonnay es baja, y $70 000 si planta únicamente uva Chardonnay y la demanda de vino Chardonnay es alta. Si únicamente planta uva Riesling, la proyección del ingreso anual es de $25 000 si la demanda de vino Riesling es baja y $45 000 si la demanda de vino Riesling es alta. Si Seneca planta ambos tipos de uva, las proyecciones de ganancias anuales son las que se muestran en la tabla siguiente.

	Demanda de Riesling	
Demanda de Chardonnay	**Baja**	**Alta**
Baja	$22 000	$40 000
Alta	$26 000	$60 000

a. ¿Cuál es la decisión a tomar, cuál es el evento aleatorio y cuál es la consecuencia?
b. Elabore un árbol de decisión.
c. Emplee el método del valor esperado para recomendar qué alternativa debe tomar Seneca para maximizar la ganancia anual.
d. Suponga que a la dirección le interesan las probabilidades estimadas para el caso de que la demanda de vino Chardonnay sea alta. Algunos suponen que en este caso la demanda de Riesling también será alta. Suponga que la probabilidad de que la demanda de Chardonnay sea alta y que la demanda de Riesling sea baja es 0.05 y que la probabilidad de que la demanda de Chardonnay sea alta y la demanda de Riesling también lo sea es 0.40. ¿Cómo modifica esto la decisión recomendada? Suponga que las probabilidades de que la demanda de Chardonnay sea baja siguen siendo 0.05 y 0.50.
e. Otros miembros del equipo directivo esperan que el mercado de Chardonnay se sature en algún momento del futuro haciendo bajar los precios. Suponga que las proyecciones de ganancia anual caigan a $50 000 si la demanda de Chardonnay es alta y sólo se siembran uvas Chardonnay. Con las estimaciones de probabilidades iniciales, determine cómo afecta este cambio a la decisión óptima.

7. El consejo municipal de Lake Placid ha decidido construir un nuevo centro comunitario que será usado para convenciones, conciertos y otros eventos públicos, pero existen controversias respecto a su tamaño. Muchos de los ciudadanos influyentes desean que sea grande para que sirva de escaparate para la zona. Pero el alcalde cree que si la demanda no lo justifica, la comunidad perderá una gran cantidad de dinero. Para facilitar la decisión, el consejo municipal ha reducido las alternativas de construcción a tres tamaños: pequeño, mediano y grande. Todo mundo coincide en que el factor relevante para elegir el tamaño es la cantidad de personas que usarán estas nuevas instalaciones. Un asesor en planeación regional proporciona estimaciones de la demanda en tres escenarios: en el peor de los casos, en el caso base y en el mejor de los casos. El peor de los casos corresponde a la situación en que el turismo baje significativamente; el caso base corresponde a la situación en que Lake Placid siga atrayendo la misma cantidad de visitantes que hasta ahora, y el mejor de los casos corresponde a un aumento significativo del turismo. El asesor ha

proporcionado las siguientes probabilidades estimadas, 0.10, 0.60 y 0.30 para el peor de los casos, el caso base y el mejor de los casos, respectivamente.

El consejo municipal ha sugerido el flujo de caja neto en un horizonte a cinco años como criterio para decidir cuál es el tamaño adecuado. Un asesor elaboró las proyecciones siguientes de flujo de caja neto (en miles de dólares) a un horizonte de cinco años. Todos los costos, incluyendo los honorarios del asesor, están incluidos.

	Escenario de demanda		
Tamaño del centro	**Peor caso**	**Caso base**	**Mejor caso**
Pequeño	400	500	660
Mediano	−250	650	800
Grande	−400	580	990

a. ¿Cuál es la decisión que deberá tomar Lake Placid con el método del valor esperado?
b. Calcule el valor esperado de la información perfecta. ¿Cree usted que valdría la pena tratar de obtener más información acerca de qué escenario tiene más posibilidades de presentarse?
c. Suponga que la probabilidad del escenario del peor de los casos aumentara a 0.2, la probabilidad del escenario del caso base disminuyera 0.5 y la probabilidad del escenario del mejor caso permaneciera igual. ¿Estos cambios tendrían algún efecto en la recomendación para la toma de decisión?
d. El asesor sugiere que un gasto de $150 000 en una campaña promocional a lo largo del horizonte de planeación reduciría a cero la probabilidad del escenario del peor caso. Si se espera que esta campaña también aumente la probabilidad del escenario del mejor de los casos a 0.4, ¿es una buena inversión?



Análisis de decisión con información muestral

Al aplicar el método del valor esperado, se mostró cómo la información de la probabilidad de los estados afecta al valor esperado y, por tanto, a la recomendación sobre la decisión. Suele ocurrir que quien debe tomar la decisión cuente con evaluaciones de probabilidad preliminar o **probabilidad previa** para los estados que son los mejores valores de probabilidad de que se dispone en ese momento. Sin embargo, para tomar la mejor decisión posible, la persona que tomará la decisión suele tratar de obtener más información acerca de los estados. Esta nueva información sirve para revisar o actualizar las probabilidades previas, de modo que la decisión final se sustente en probabilidades más certeras de los estados. Lo más frecuente es que se obtenga más información mediante experimentos diseñados para proporcionar **información muestral** acerca de los estados. El muestreo de materia prima, la prueba de productos y los estudios de investigación de mercado son ejemplos de experimentos (o estudios) que permiten a los directivos modificar o actualizar las probabilidades de los estados. A estas probabilidades actualizadas se les llama **probabilidades posteriores**.

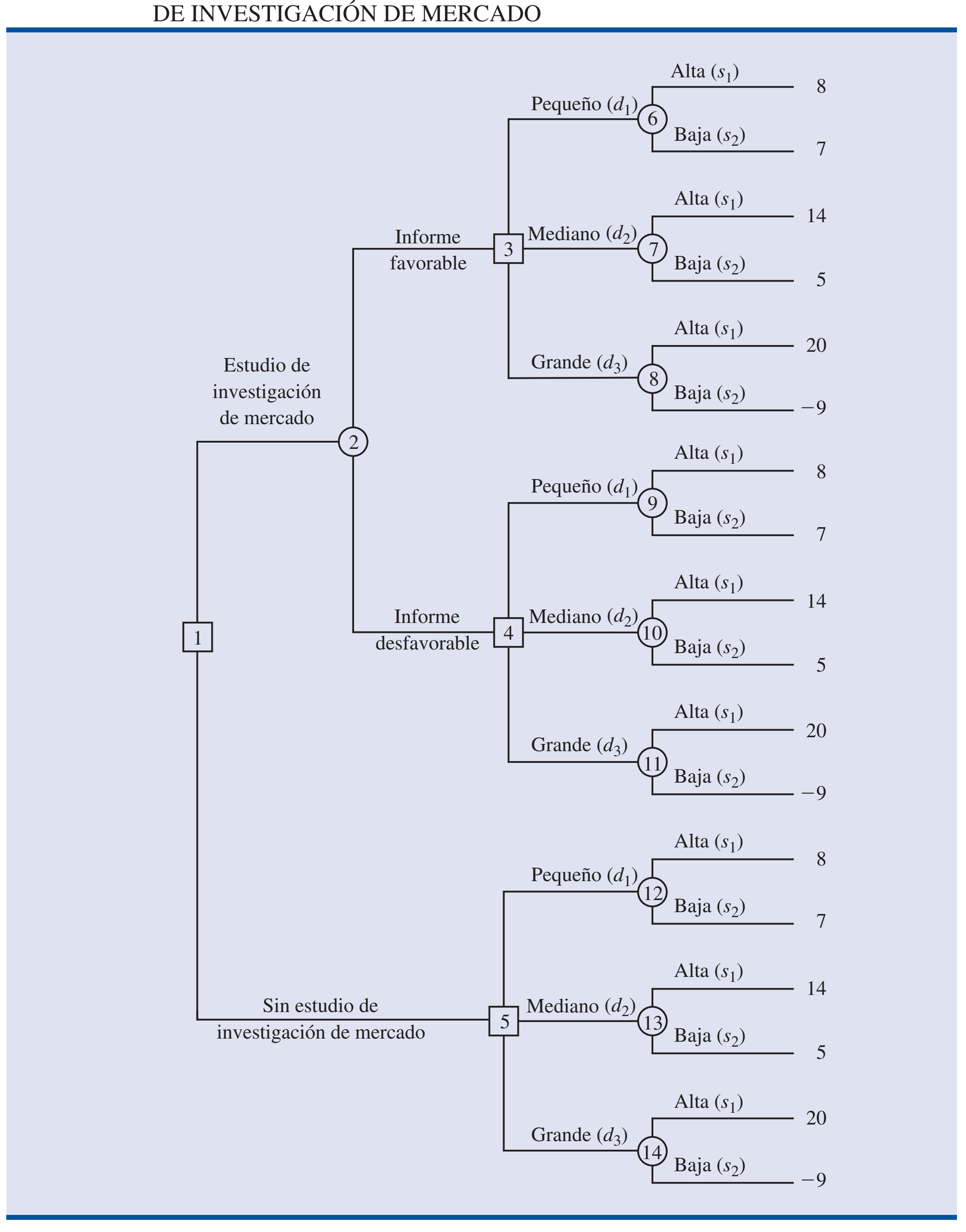
De regreso al ejemplo de PDC, suponga que el director considera la posibilidad de hacer un estudio de investigación de mercado de seis meses de duración para conocer mejor la aceptación potencial en el mercado del proyecto de los condominios de PDC. El director prevé que el estudio de la investigación de mercado proporcionará uno de los dos siguientes resultados:

1. Informe favorable: Una cantidad significativa de las personas entrevistadas expresó interés por comprar un condominio de PDC.
2. Informe desfavorable: Muy pocas de las personas entrevistadas expresaron interés por comprar un condominio de PDC.

Árbol de decisión

En la figura 21.4 se presenta el árbol de decisión para el problema de PDC con información muestral, en el que se observa la secuencia lógica de las decisiones y de los eventos aleatorios. Primero, el director de PDC debe decidir si llevar a cabo el estudio de la investigación de mercado. Si se realiza este estudio, el director de PDC debe estar preparado para tomar una decisión acerca del tamaño del complejo de condominios en caso de que el informe del estudio de la investigación de mercado sea favorable y, tal vez, otra decisión distinta acerca del tamaño del complejo si el informe de la investigación de mercado es desfavorable.

FIGURA 21.4 ÁRBOL DE DECISIÓN DE PDC QUE COMPRENDE EL ESTUDIO DE INVESTIGACIÓN DE MERCADO



En la figura 21.4 los cuadrados indican nodos de decisión y los círculos indican nodos aleatorios. En cada nodo de decisión, la rama que se siga depende de la decisión que se tome. En cada nodo aleatorio, la rama que se siga depende de la probabilidad. Por ejemplo, el nodo de decisión 1 indica que PDC debe decidir si realiza el estudio de investigación de mercado. Si realiza el estudio de investigación de mercado, el nodo aleatorio 2 indica que las ramas del informe favorable y del informe desfavorable no se encuentran bajo control de PDC y estarán determinadas por la casualidad. El nodo 3 es de decisión e indica que si el informe de la investigación de mercado es favorable, PDC debe decidir si el complejo que construya será pequeño, mediano o grande. El nodo 4 es un nodo de decisión que indica que si el reporte de la investigación de mercado es desfavorable, PDC debe decidir si el complejo que construya será pequeño, mediano o grande. El nodo 5 es un nodo de decisión que indica que si PDC no realiza la investigación de mercado debe decidir si el complejo que construya será pequeño, mediano o grande. Los nodos 6 a 14 son aleatorios e indican que las ramas de los estados alta demanda o baja demanda serán determinadas por la casualidad.

En la sección 21.4 se explica cómo obtener estas probabilidades.

Para el análisis de un árbol de decisión y para la elección de una estrategia óptima es necesario que se conozcan todas las probabilidades de rama de todos los nodos aleatorios. PDC tiene las siguientes probabilidades de rama.

Si se realiza el estudio de investigación de mercado

$$P(\text{Informe favorable}) = P(F) = 0.77$$
$$P(\text{Informe desfavorable}) = P(D) = 0.23$$

Si el informe de la investigación de mercado es favorable

$$P(\text{Demanda alta dado un informe favorable}) = P(s_1|\text{F}) = 0.94$$
$$P(\text{Demanda baja dado un informe favorable}) = P(s_2|\text{F}) = 0.06$$

Si el informe de la investigación de mercado es desfavorable

$$P(\text{Demanda alta dado un informe desfavorable}) = P(s_1|\text{D}) = 0.35$$
$$P(\text{Demanda baja dado un informe desfavorable}) = P(s_2|\text{D}) = 0.65$$

Si no se realiza la investigación de mercado, se pueden emplear las probabilidades previas.

$$P(\text{Demanda alta}) = P(s_1) = 0.80$$
$$P(\text{Demanda baja}) = P(s_2) = 0.20$$

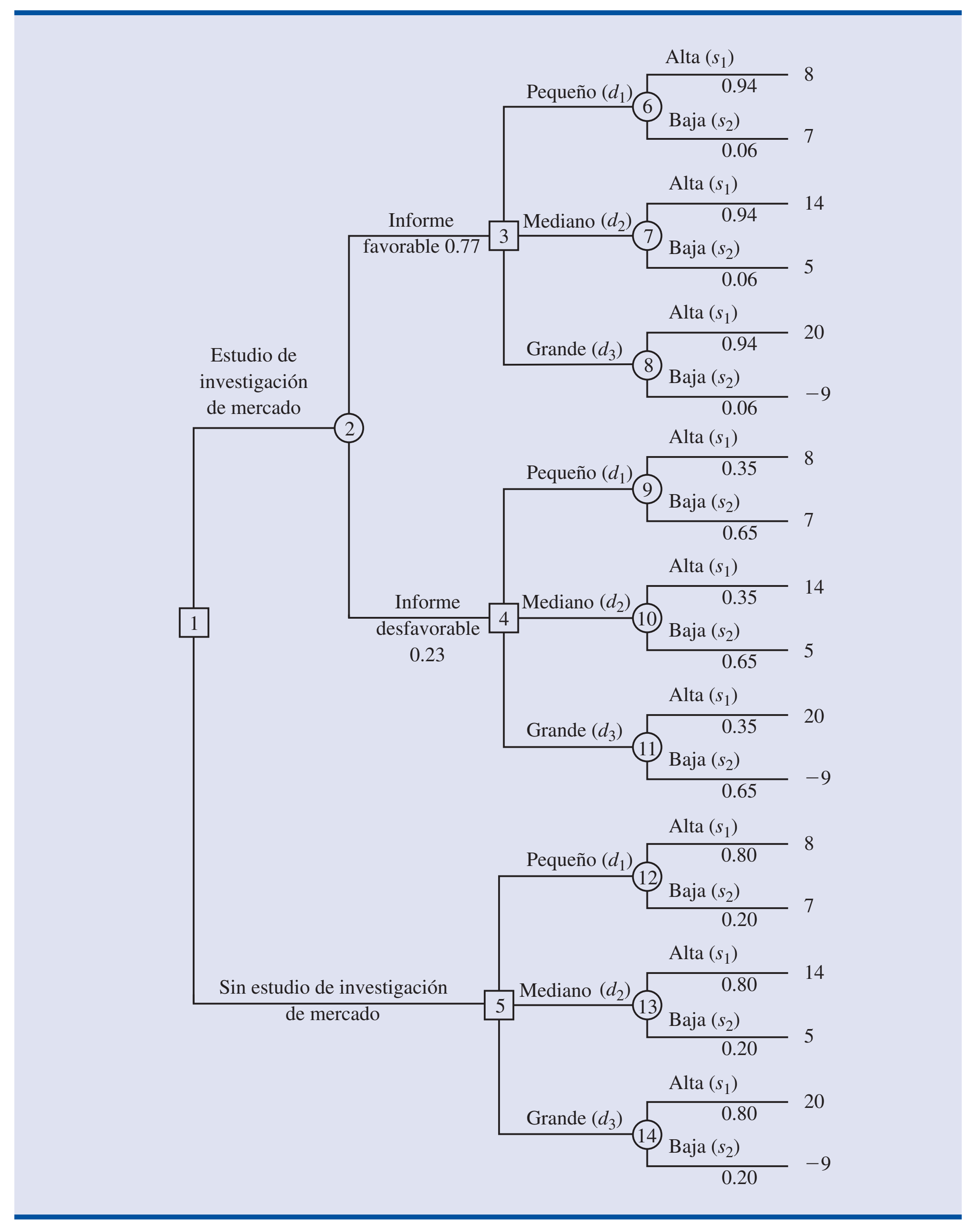
En el árbol de decisión de la figura 21.5, sobre el árbol de decisión, se muestran las probabilidades de rama.

Estrategia de decisión

Una **estrategia de decisión** es una secuencia de decisiones y resultados aleatorios, donde las decisiones que se toman dependen de los resultados, por conocer, de los eventos aleatorios. El método que se emplea para determinar la estrategia óptima de decisión se basa en recorrer el árbol de decisión en sentido regresivo, de atrás para adelante, debe seguir los pasos que se indican a continuación:

1. En los nodos aleatorios se calcula su valor esperado al multiplicar la recompensa que aparece al final de cada rama por la correspondiente probabilidad de la rama.
2. En los nodos de decisión, se elige la rama de decisión que conduzca al mayor valor esperado. Este valor esperado será el valor esperado del nodo de decisión.

FIGURA 21.5 ÁRBOL DE DECISIÓN DE PDC QUE MUESTRA LAS PROBABILIDADES DE RAMA



Al comenzar el recorrido regresivo con el cálculo de los valores esperados de los nodos 6 a 14 se obtienen los resultados siguientes:

$$\begin{aligned}
\text{VE(Nodo 6)} &= 0.94(8) + 0.06(7) = 7.94\\
\text{VE(Nodo 7)} &= 0.94(14) + 0.06(5) = 13.46\\
\text{VE(Nodo 8)} &= 0.94(20) + 0.06(-9) = 18.26\\
\text{VE(Nodo 9)} &= 0.35(8) + 0.65(7) = 7.35\\
\text{VE(Nodo 10)} &= 0.35(14) + 0.65(5) = 8.15\\
\text{VE(Nodo 11)} &= 0.35(20) + 0.65(-9) = 1.15\\
\text{VE(Nodo 12)} &= 0.80(8) + 0.20(7) = 7.80\\
\text{VE(Nodo 13)} &= 0.80(14) + 0.20(5) = 12.20\\
\text{VE(Nodo 14)} &= 0.80(20) + 0.20(-9) = 14.20
\end{aligned}$$

En la figura 21.6 se presenta el árbol de decisión reducido, una vez calculados los valores esperados de estos nodos aleatorios.

Después se continúa con los nodos 3, 4, y 5. En cada uno de estos nodos se elige la rama de la alternativa de decisión que conduzca al mayor valor esperado. Por ejemplo, en el nodo 3 se puede elegir entre las ramas del complejo pequeño para la que VE(Nodo 6) = 7.49, la del complejo mediano para la que VE(Nodo 7) = 13.46 y la del complejo grande para la que VE(Nodo 8) = 18.26. Se elegirá la rama de la alternativa de decisión del complejo grande y el valor esperado del nodo 3 será, VE(Nodo 3) = 18.26.

En el nodo 4 hay que elegir, entre los nodos 9, 10 y 11, el que tenga el mayor valor esperado. La mejor alternativa de decisión es la rama del complejo mediano, con la que se obtiene VE(Nodo 4) = 8.15. En el nodo 5 la elección es entre los nodos 12, 13 y 14, el que tenga el mayor valor esperado. La mejor alternativa de decisión es la rama con la que se obtiene VE(Nodo 5) = 14.20. En la figura 21.7 se presenta el árbol de decisión reducido una vez que en los nodos 3, 4, y 5 se han elegido, o tomado, las mejores decisiones.

Ahora se puede calcular el valor esperado del nodo 2, de la manera siguiente:

$$\begin{aligned}
\text{VE(Nodo 2)} &= 0.77\text{VE(Nodo 3)} + 0.23\text{VE(Nodo 4)}\\
&= 0.77(18.26) + 0.23(8.15) = 15.93
\end{aligned}$$

Con estos cálculos el árbol de decisión se reduce a un árbol que tiene únicamente dos ramas de decisión que salen del nodo 1 (véase la figura 21.8).

Por último, en el nodo 1 se puede llegar a una decisión al elegir entre los nodos 2 y 5, el que tenga el mayor valor esperado. Esto lleva a la alternativa de decisión de realizar el estudio de investigación de mercado, con el cual se obtendrá un valor esperado de 15.93.

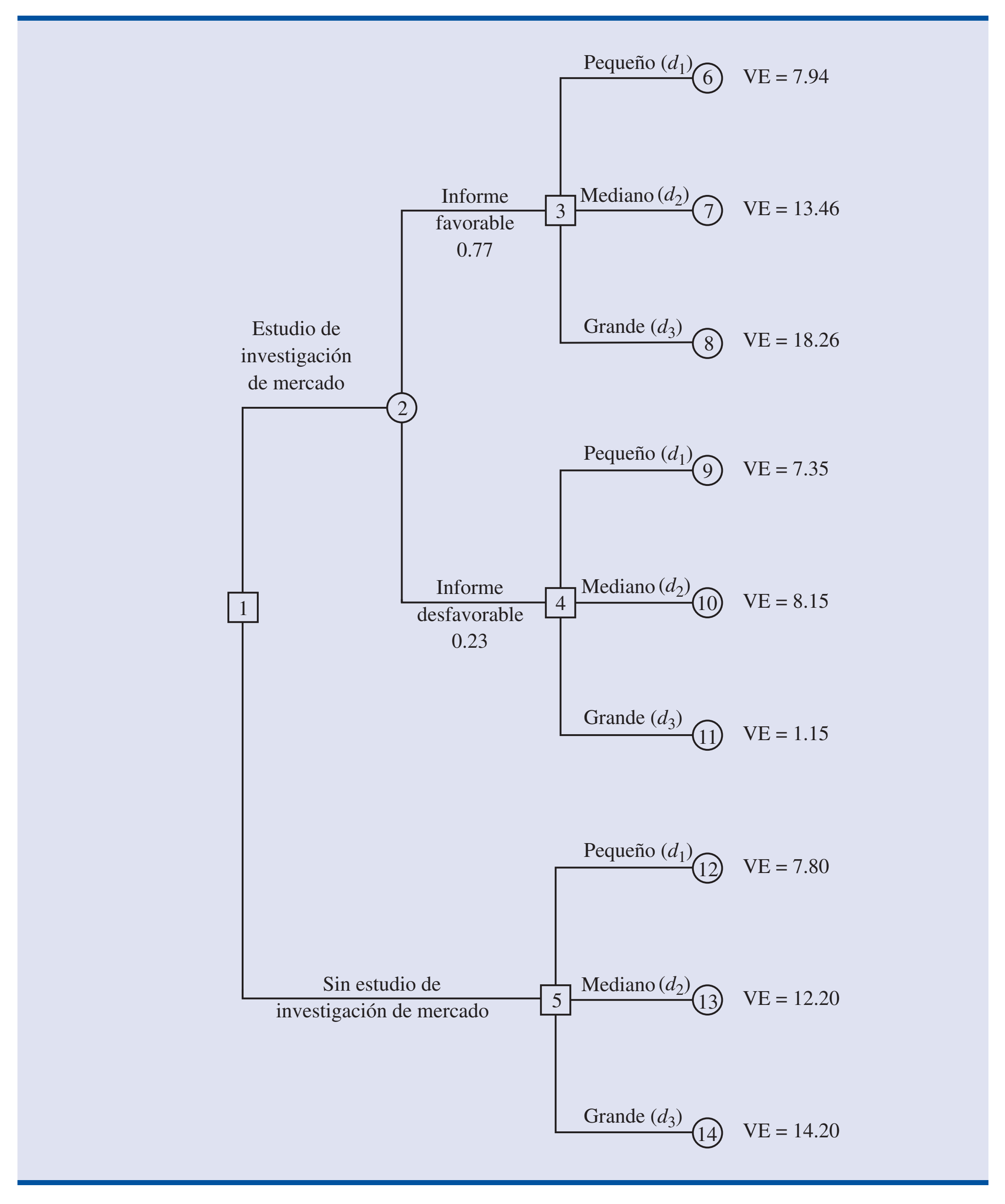
Para PDC, la decisión óptima es realizar el estudio de investigación de mercado y después seguir la siguiente estrategia de decisión:

Si la investigación de mercado es favorable, construir el complejo grande.

Si la investigación de mercado es desfavorable, construir el complejo mediano.

Este análisis del árbol de decisión de PDC ilustra los métodos que pueden usarse para analizar problemas secuenciales de decisiones más complejos. Primero, se dibuja el árbol de decisión que consta de nodos de decisión, nodos aleatorios y ramas que describen el carácter secuencial del problema. Se determinan las probabilidades de todos los resultados aleatorios. Después procediendo en forma regresiva a través del árbol de decisión se calcula el valor esperado de todos los nodos aleatorios y en cada uno de los nodos de decisión se elige la rama que conduzca a la mejor decisión. La secuencia de ramas de decisiones óptimas determina la estrategia de la decisión óptima para el problema.

FIGURA 21.6 ÁRBOL DE DECISIÓN DE PDC UNA VEZ CALCULADOS LOS VALORES ESPERADOS DE LOS NODOS DE DECISIÓN 6 A 14

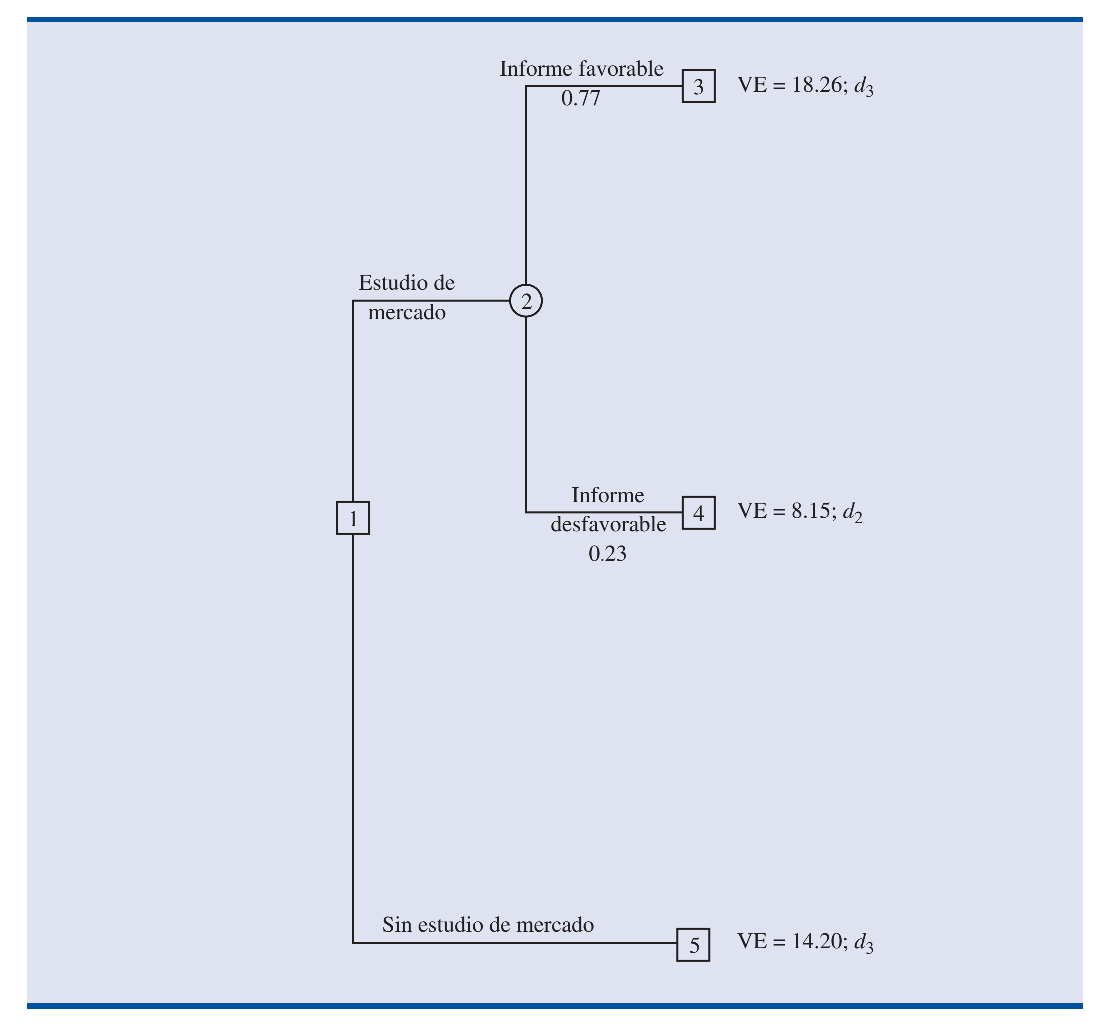


Valor esperado de la información muestral

El VEIM = $1.73 millones sugiere que PDC deberá estar dispuesto a pagar hasta $1.73 millones por la realización del estudio de la investigación de mercado.

En el problema de PDC, la información muestral que se usa para determinar la estrategia óptima de decisión es el estudio de la investigación de mercado. El valor esperado del estudio de la investigación de mercado es $15.93. En la sección 21.3 se mostró que si *no* se realiza la investigación de mercado, el mejor valor esperado es $14.20. Por tanto, se concluye que la diferencia $15.93 − 14.20 = $1.73 es el **valor esperado de la información muestral (VEIM)**. En otras

FIGURA 21.7 ÁRBOL DE DECISIÓN DE PDC, UNA VEZ ELEGIDAS LAS MEJORES DECISIONES EN LOS NODOS 3, 4 Y 5



palabras, realizar el estudio de la investigación de mercado agrega \$1.73 millones al valor esperado de PDC. En general, el valor esperado de la información muestral es el siguiente:

VALOR ESPERADO DE LA INFORMACIÓN MUESTRAL

$$\text{VEIM} = |\text{VEcIM} - \text{VEsIM}| \qquad \textbf{(21.5)}$$

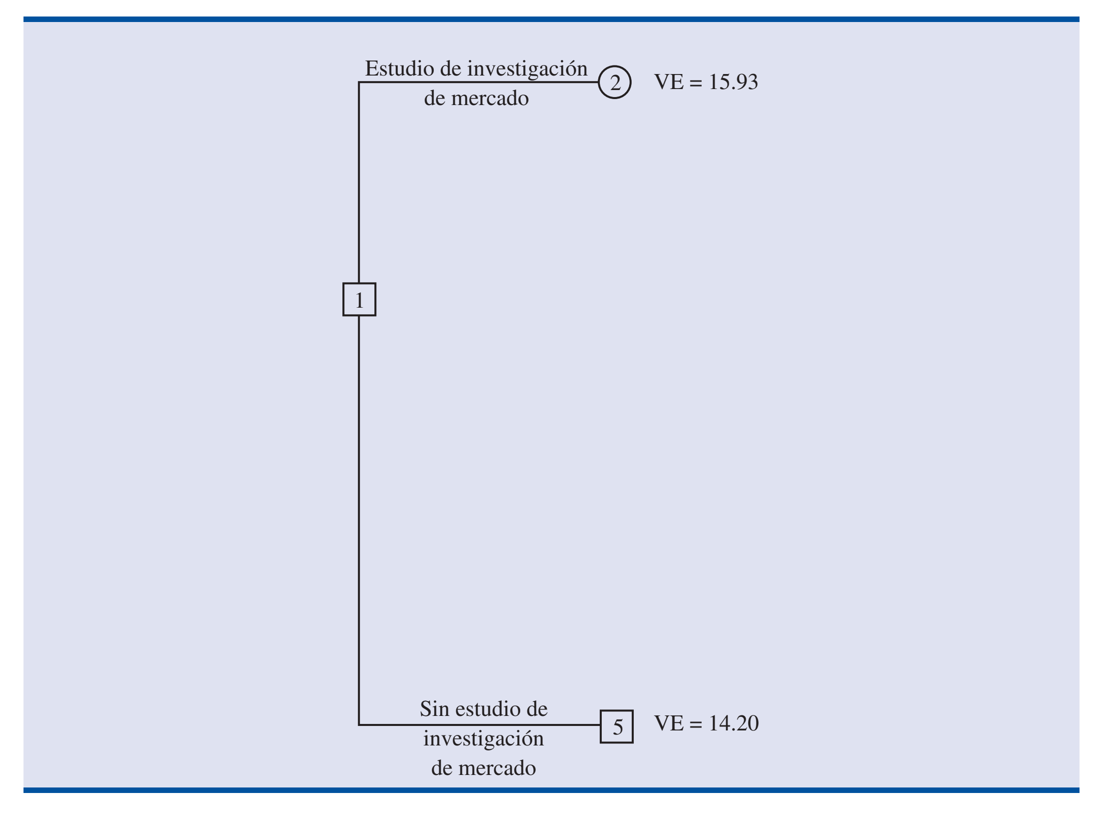
donde

VEIM = valor esperado de la información muestral
VEcIM = valor esperado *con* información muestral acerca del estado
VEsIM = valor esperado *sin* información muestral acerca del estado

Note el papel del valor absoluto en la ecuación (21.5). En problemas de minimización, el valor esperado con información muestral es siempre menor que el valor esperado sin información



FIGURA 21.8 ÁRBOL DE DECISIÓN DE PDC REDUCIDO A DOS RAMAS DE DECISIÓN



muestral. En ese caso VEIM es la magnitud de la diferencia entre VEcIM y VEsIM; de esta manera, al tomar el valor absoluto de la diferencia, como se muestra en la ecuación (21.5), se pueden tratar tanto los casos de maximización como los de minimización con una sola ecuación.

Ejercicios

Métodos



8. Se considerará una variación del árbol de decisión de PDC mostrado en la figura 21.5. Lo primero que tiene que decidir la empresa es si llevar a cabo, o no, el estudio de la investigación de mercado. Si se realiza el estudio de la investigación de mercado, los resultados pueden ser favorables (F) o desfavorables (D). Ahora suponga que sólo se tienen dos alternativas de decisión, d_1 y d_2, y dos estados, s_1 y s_2. En la tabla de recompensas siguiente se muestran las ganancias:

	Estado	
Alternativa de decisión	s_1	s_2
d_1	100	300
d_2	400	200

a. Presente el árbol de decisión.
b. A partir de las probabilidades siguientes proporcione la estrategia óptima de decisión.

$$P(F) = 0.56 \quad P(s_1 \mid F) = 0.57 \quad P(s_1 \mid D) = 0.18 \quad P(s_1) = 0.40$$
$$P(D) = 0.44 \quad P(s_2 \mid F) = 0.43 \quad P(s_2 \mid D) = 0.82 \quad P(s_2) = 0.60$$

Aplicaciones

9. Un inversionista de bienes raíces tiene la oportunidad de comprar un terreno en una zona que actualmente es residencial. Si en el lapso de un año, la junta de administración local aprueba la solicitud de modificar el uso de esta propiedad a propiedad comercial, el inversionista podrá rentar el terreno a una tienda de descuento que desea abrir una sucursal en ese lugar. Pero si esta modificación no es aprobada, el inversionista tendría que vender la propiedad y tener una pérdida. En la siguiente tabla de recompensa se presentan las ganancias posibles (en miles de dólares).

	Estado	
Alternativas de decisión	**Aprobación** s_1	**No aprobación** s_2
Comprar, d_1	600	−200
No comprar, d_2	0	0

a. Si la probabilidad de que se acepte la modificación es 0.5, ¿cuál es la decisión que se recomienda?
b. Al comprar el terreno el inversionista accede a una opción, la cual le concede el derecho de comprar el terreno dentro de los próximos tres meses, tiempo en el que tendrá oportunidad de obtener más información acerca de la resistencia de los habitantes de la zona a la modificación solicitada. Las probabilidades son las siguientes:

Sea $M =$ mucha resistencia a la modificación
$P =$ poca resistencia a la modificación

$$P(M) = 0.55 \quad P(s_1 \mid M) = 0.18 \quad P(s_2 \mid M) = 0.82$$
$$P(P) = 0.45 \quad P(s_1 \mid P) = 0.89 \quad P(s_2 \mid P) = 0.11$$

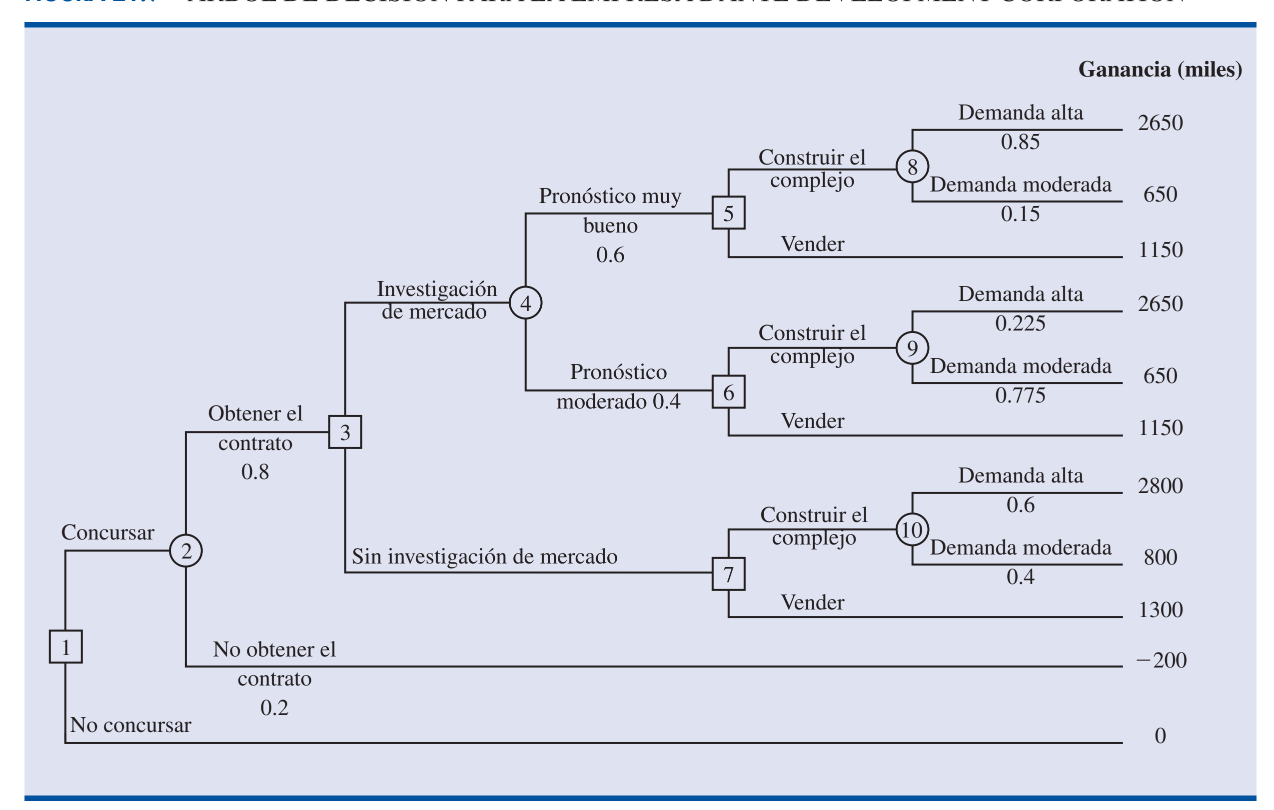
¿Cuál es la estrategia optima de decisión si el inversionista aprovecha el periodo que le da la opción, para obtener más información acerca de la resistencia de los habitantes de la zona, antes de tomar una decisión sobre la compra?
c. Si para adquirir la opción, el inversionista tiene que gastar $10 000 más, ¿deberá adquirir la opción? ¿Por qué sí o por qué no? ¿Cuál será la máxima cantidad que el inversionista estará dispuesto a pagar por la opción?

10. La empresa Dante Development Corporation considera la posibilidad de concursar por un contrato para la construcción de un nuevo complejo de oficinas. En la figura 21.9 se presenta el árbol de decisión elaborado por uno de los analistas de Dante. En el nodo 1 la empresa tiene que decidir si concursa o no por el contrato. Prepararse para el concurso cuesta $200 000. La rama superior del nodo 2 indica que, si la empresa concursa, la probabilidad de ganar el contrato es 0.8. Si la empresa gana el contrato, tendrá que pagar $2 000 000 para convertirse en socio del proyecto. El nodo 3 muestra que, entonces, la empresa, antes de empezar con la construcción, tendrá que considerar la posibilidad de realizar un estudio de investigación de mercado para pronosticar la demanda que tendrán las oficinas. El costo de este estudio es de $150 000. El nodo 4 es aleatorio y señala los posibles resultados del estudio de la investigación de mercado.

Los nodos 5, 6 y 7 son todos similares entre sí, ya que todos son nodos de decisión en los que Dante tiene que decidir si construir las oficinas o vender el derecho sobre el proyecto a otra empresa. Si se decide a construir el complejo, tendrá un ingreso de $5 000 000 si la demanda es alta, y $3 000 000, si la demanda es moderada. Si Dante decide vender sus derechos sobre el proyecto a otra empresa, se estima que el ingreso por la venta será $3 500 000. Las probabilidades en los nodos 4, 8 y 9 están basadas en los resultados del estudio de la investigación de mercado.

a. Verifique las proyecciones de ganancias que se presentan al final de las ramas del árbol de decisión, calcule las recompensas de $2 650 000 y $650 000 de los dos primeros resultados.
b. ¿Cuál es la estrategia óptima de decisión para Dante, y cuál es la ganancia esperada en este proyecto?
c. ¿De cuánto tendrá que ser el costo del estudio de la investigación de mercado para que Dante se decida a realizar el estudio?

FIGURA 21.9 ÁRBOL DE DECISIÓN PARA LA EMPRESA DANTE DEVELOPMENT CORPORATION



11. La empresa Hale's TV Productions considera producir un programa piloto de una serie de televisión que espera vender a una cadena de televisión. Puede ser que la cadena rechace la serie, pero también que decida comprar los derechos de la serie por uno o dos años. En este momento, Hale tiene que producir la muestra y esperar a la decisión de la cadena de televisión o transferirle los derechos sobre el piloto y sobre la serie a un competidor por $100 000. En la tabla siguiente se muestran las alternativas de decisión y las ganancias de Hale.

	Estado		
Alternativas de decisión	**Rechazo, s_1**	**1 año, s_2**	**2 años, s_3**
Producir la muestra, d_1	−100	50	150
Vender a la competencia, d_2	100	100	100

Las probabilidades de los estados son $P(s_1) = 0.2$, $P(s_2) = 0.3$ y $P(s_3) = 0.5$. Mediante un pago de $5 000, una agencia puede revisar los planes de la serie e indicar las posibilidades de que la cadena de televisión tenga una reacción favorable. La revisión de la agencia puede dar resultados favorables (F) o desfavorables (D); suponga que las probabilidades sean las siguientes.

$$P(F) = 0.69 \quad P(s_1 \mid F) = 0.90 \quad P(s_1 \mid D) = 0.45$$
$$P(D) = 0.31 \quad P(s_2 \mid F) = 0.26 \quad P(s_2 \mid D) = 0.39$$
$$P(s_3 \mid F) = 0.65 \quad P(s_3 \mid D) = 0.16$$

a. Construya un árbol de decisión para este problema.
b. ¿Cuál es la decisión recomendada si no se usa la opinión de la agencia? ¿Cuál es el valor esperado?

c. ¿Cuál es el valor esperado de la información perfecta?
d. ¿Cuál es la estrategia óptima de decisión para Hale si usa la información de la agencia?
e. ¿Cuál es el valor esperado de la información de la agencia?
f. ¿La información de la agencia vale los $5 000 que hay que pagarle? ¿Cuál es la cantidad máxima que Hale estará dispuesta a pagar por la información?
g. ¿Cuál es la decisión recomendada?

12. Martin's Service Station está considerando participar, el próximo invierno, en el negocio de quitar la nieve. Martin puede comprar una cuchilla aditamento para su camión pick-up o un camión quitanieves para trabajo pesado. Después de analizar la situación, Martin encuentra que cualquier alternativa será una inversión rentable si hay nevadas fuertes. Si las nevadas son moderadas, la rentabilidad puede ser pequeña y si las nevadas son muy ligeras, puede tener pérdidas. En la tabla siguiente se muestran las ganancias y las pérdidas

	Estado		
Alternativas de decisión	**Fuerte, s_1**	**Moderada, s_2**	**Ligera, s_3**
Aditamento, d_1	3500	1000	−1500
Quitanieves, d_2	7000	2000	−9000

Las probabilidades de los estados son $P(s_1) = 0.4$, $P(s_2) = 0.3$ $P(s_3) = 0.3$. Suponga que Martin decide esperar hasta septiembre antes de tomar una decisión. Las probabilidades estimadas de que en septiembre haga un frío normal (N) o un frío inesperado (I) son las siguientes:

$$P(N) = 0.8 \quad P(s_1 \mid N) = 0.35 \quad P(s_1 \mid U) = 0.62$$
$$P(U) = 0.2 \quad P(s_2 \mid N) = 0.30 \quad P(s_2 \mid U) = 0.31$$
$$P(s_3 \mid N) = 0.35 \quad P(s_3 \mid U) = 0.07$$

a. Construya un árbol de decisión para este problema.
b. ¿Cuál es la decisión recomendada si Martin no espera hasta septiembre? ¿Cuál es el valor esperado?
c. ¿Cuál es el valor esperado de la información perfecta?
d. ¿Cuál es la estrategia óptima de decisión si no se toma la decisión sino hasta que se haya determinado el clima en septiembre? ¿Cuál es el valor esperado de esta estrategia?

13. La tienda departamental Lawson tiene que decidir si compra un producto estacional, el cual puede tener una demanda alta, moderada o baja. Lawson puede ordenar 1, 2, o 3 lotes, antes de la estación, pero después ya no podrá comprar otro lote. Las proyecciones de las ganancias (en miles de dólares) son las que se muestran a continuación.

	Estado		
Alternativas de decisión	**Demanda alta s_1**	**Demanda moderada s_2**	**Demanda baja s_3**
Ordenar 1 lote, d_1	60	60	50
Ordenar 2 lotes, d_2	80	80	30
Ordenar 3 lotes, d_3	100	70	10

a. Si las probabilidades previas de los tres estados son 0.3, 0.3 y 0.4, respectivamente, ¿cuál es la cantidad de lotes recomendada?
b. En todas las reuniones preestacionales de ventas, el vicepresidente de la empresa da su opinión respecto a la demanda potencial de los productos. Debido al carácter entusiasta y optimista del vicepresidente, los pronósticos de las condiciones del mercado siempre han sido

o "excelentes" (E) o "muy buenas" (M). Las probabilidades son las siguientes. ¿Cuál es la estrategia óptima de decisión?

$$\begin{array}{lll} P(E) = 0.7 & P(s_1 \mid E) = 0.34 & P(s_1 \mid M) = 0.20 \\ P(M) = 0.3 & P(s_2 \mid E) = 0.32 & P(s_2 \mid M) = 0.26 \\ & P(s_3 \mid E) = 0.34 & P(s_3 \mid M) = 0.54 \end{array}$$

c. Calcule el VEIP y el VEIM. Analice si la empresa debería consultar a un experto que le proporcionara un pronóstico independiente de la situación del mercado respecto al producto.



21.4 Cálculo de las probabilidades de rama mediante el teorema de Bayes

En la sección 21.3 las probabilidades de rama de los nodos aleatorios del árbol de decisión fueron especificadas en la descripción del problema. Para determinar estas probabilidades no fue necesario realizar ningún cálculo. En esta sección se muestra cómo usar el **teorema de Bayes**, ya visto en el capítulo 4, para calcular las probabilidades de rama de un árbol de decisión.

En la figura 21.10 se presenta nuevamente el árbol de decisión de PDC. Sea

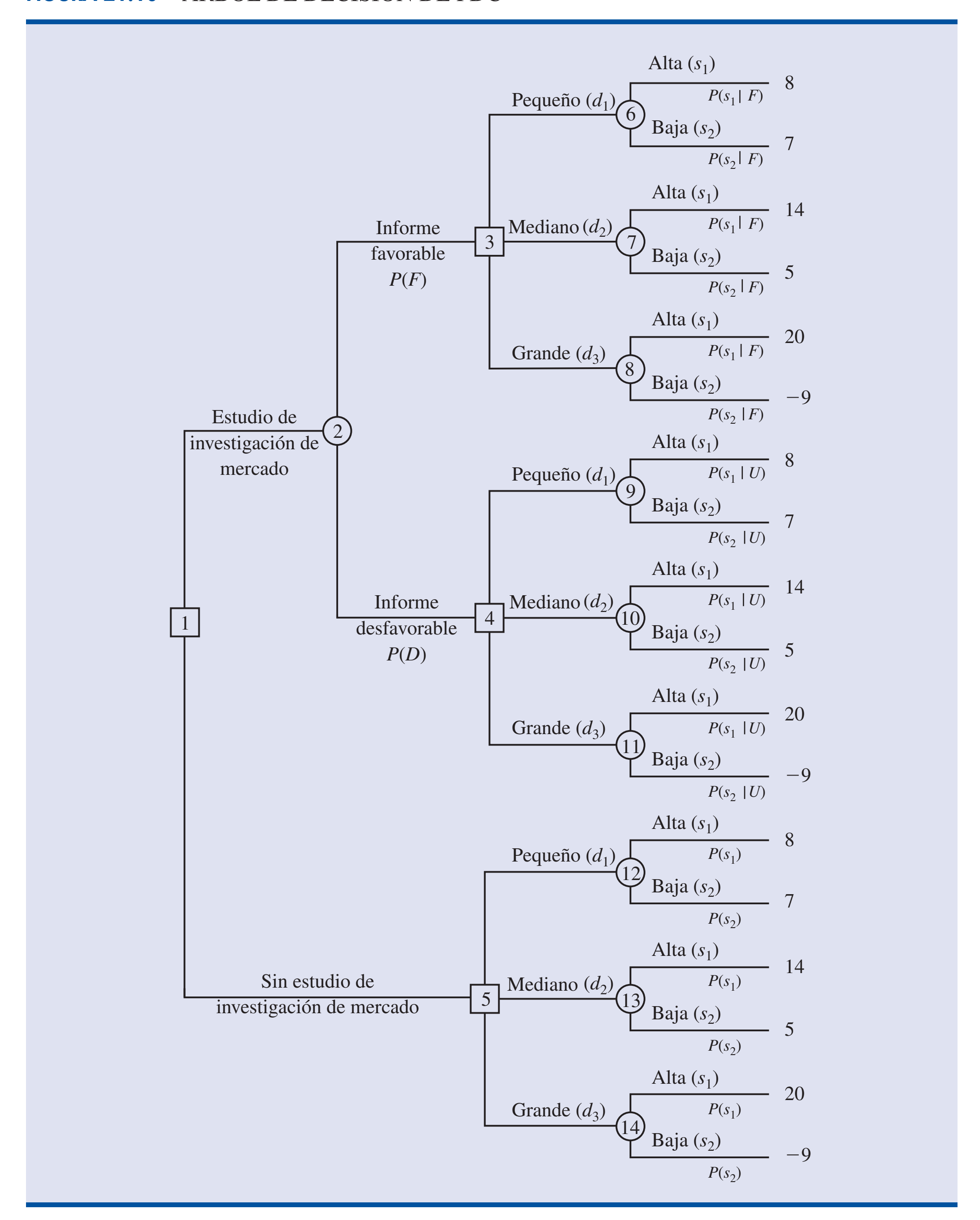
F = Informe favorable de la investigación de mercado
D = Informe desfavorable de la investigación de mercado
s_1 = Demanda alta (estado 1)
s_2 = Demanda baja (estado 2)

En el nodo aleatorio 2, se necesitan conocer las probabilidades de rama $P(F)$ y $P(D)$. En los nodos aleatorios 6, 7 y 8 se necesitan conocer las probabilidades de rama, $P(s_1|F)$ probabilidad del estado 1 dado un informe favorable de la investigación de mercado y $P(s_2|F)$ probabilidad del estado 2 dado un informe favorable de la investigación de mercado. A $P(s_1|F)$ y $P(s_2|F)$ se les conoce como *probabilidades posteriores* debido a que son probabilidades condicionales basadas en el resultado de la información muestral. En los nodos 9, 10 y 11 se necesitan conocer las probabilidades de rama $P(s_1|D)$ y $P(s_2|D)$; observe que éstas también son probabilidades posteriores que denotan las probabilidades de dos estados *dado* que el informe de la investigación de mercado es desfavorable. Por último, en los nodos 12, 13 y 14 se necesitan las probabilidades de los estados $P(s_1)$ y $P(s_2)$ si no se realiza el estudio de la investigación de mercado.

Para calcular las probabilidades, se necesita saber cuáles son los valores que PDC da a las probabilidades de los estados $P(s_1)$ y $P(s_2)$, que son las probabilidades previas, como ya se vio antes. Además, se necesita conocer la **probabilidad condicional** de los resultados de la investigación de mercado (información muestral) *dado* cada uno de los estados. Por ejemplo, se necesita conocer la probabilidad condicional de un informe favorable de la investigación de mercado dado que existe una alta demanda por el proyecto de PDC; observe que esta probabilidad condicional de F dado s_1 se expresa $P(F|s_1)$. Para calcular estas probabilidades, se necesitan las probabilidades condicionales de todos los resultados muestrales dados cada uno de los estados, es decir, $P(F|s_1)$, $P(F|s_2)$, $P(D|s_1)$ y $P(D|s_2)$. En el caso del problema de PDC se supone que se cuenta con las estimaciones de las siguientes probabilidades condicionales.

	Investigación de mercado	
Estado	**Favorable, *F***	**Desfavorable, *D***
Demanda alta, s_1	$P(F \mid s_1) = 0.90$	$P(U \mid s_1) = 0.10$
Demanda baja, s_2	$P(F \mid s_2) = 0.25$	$P(U \mid s_2) = 0.75$

FIGURA 21.10 ÁRBOL DE DECISIÓN DE PDC



Observe que las estimaciones anteriores proporcionan una confianza razonable en el estudio de la investigación de mercado. Si el verdadero estado es s_1, la probabilidad de que el informe de la investigación de mercado sea favorable es 0.90 y la probabilidad de que el informe de la investigación de mercado sea desfavorable es 0.10. Si el verdadero estado es s_2, la probabilidad de que el informe de la investigación de mercado sea favorable es 0.25 y la probabilidad de que el informe de la investigación de mercado sea desfavorable es 0.75. La razón de que la probabilidad de un potencialmente engañoso informe favorable de mercado sea 0.25 para el estado s_2 es que cuando los compradores potenciales oyen por primera vez hablar del nuevo proyecto del

TABLA 21.3 PROBABILIDADES DE RAMA PARA EL PROYECTO DEL CONDOMINIO DE PDC BASADAS EN UN REPORTE FAVORABLE DE LA INVESTIGACIÓN DE MERCADO

Estados s_j	**Probabilidades previas** $P(s_j)$	**Probabilidades condicionales** $P(F \mid s_j)$	**Probabilidades conjuntas** $P(F \cap s_j)$	**Probabilidades posteriores** $P(s_j \mid F)$
s_1	0.8	0.90	0.72	0.94
s_2	0.2	0.25	0.05	0.06
	1.0		$P(F) = 0.77$	1.00

condominio el entusiasmo puede llevarlos a exagerar su verdadero interés por los condominios. La respuesta inicialmente favorable de un comprador potencial puede convertirse rápidamente en un "no gracias" cuando se encuentra ante la situación de firmar un contrato de compra y tener que hacer un pago.

En el análisis siguiente, para calcular las probabilidades se emplea un método tabular. En la tabla 21.3 se presentan en forma resumida los cálculos para el problema de PDC cuando de la investigación de mercado se obtiene un informe favorable (F). Los pasos para elaborar esta tabla son los siguientes.

Paso 1. En la columna 1 se ingresan los estados. En la columna 2 se ingresan las *probabilidades previas* de los estados. En la columna 3 se ingresan las *probabilidades condicionales* correspondientes a un informe favorable de la investigación de mercado (F) dado cada estado.

Paso 2. En la columna 4 se calculan las **probabilidades conjuntas** al multiplicar los valores de las probabilidades previas de la columna 2 por los correspondientes valores de probabilidad condicional de la columna 3.

Paso 3. Las probabilidades conjuntas de la columna 4 se suman para obtener la probabilidad de un reporte favorable de la investigación de mercado, $P(F)$.

Paso 4. Cada probabilidad conjunta de la columna 4 se divide entre $P(F) = 0.77$ para obtener las *probabilidades posteriores,* $P(s_1|F)$ y $P(s_2|F)$.

En la tabla 21.3 se observa que la probabilidad de que el informe de la investigación de mercado sea favorable es $P(F) = 0.77$. Además, $P(s_1|F) = 0.94$ y $P(s_2|F) = 0.06$. En particular, que si el informe de la investigación de mercado es favorable, la probabilidad posterior de que la demanda por el condominio sea alta será 0.94.

El procedimiento para calcular las probabilidades en forma tabular deberá repetirse con cada uno de los resultados de la información muestral. En la tabla 21.4 se presenta el cálculo de las probabilidades de rama cuando el informe del mercado es desfavorable. Observe que la probabilidad de que el informe de la investigación de mercado sea desfavorable es $P(D) = 0.23$. Si

TABLA 21.4 PROBABILIDADES DE RAMA PARA EL PROYECTO DEL CONDOMINIO DE PDC BASADAS EN UN REPORTE DESFAVORABLE DE LA INVESTIGACIÓN DE MERCADO

Estados s_j	**Probabilidades previas** $P(s_j)$	**Probabilidades condicionales** $P(D \mid s_j)$	**Probabilidades conjuntas** $P(D \cap s_j)$	**Probabilidades posteriores** $P(s_j \mid D)$
s_1	0.8	0.10	0.08	0.35
s_2	0.2	0.75	0.15	0.65
	1.0		$P(U) = 0.23$	1.00

el informe de la investigación de mercado es desfavorable, la probabilidad posterior de que la demanda sea alta s_1, es 0.35 y la probabilidad posterior de que la demanda sea baja, s_2 es 0.65. Las probabilidades de rama de las tablas 21.3 y 21.4 se presentaron en el árbol de decisión de PDC de la figura 21.5.

En el ejercicio 14 se pide calcular probabilidades posteriores.

Lo visto en esta sección indica que existe una relación entre las probabilidades de las diferentes ramas de un árbol de decisión. No sería correcto suponer otras probabilidades previas $P(s_1)$ y $P(s_2)$ y no determinar cómo tal modificación alteraría $P(F)$ y $P(D)$, así como a las probabilidades posteriores $P(s_1|F)$, $P(s_2|F)$, $P(s_1|D)$ y $P(s_2|D)$.

Ejercicios

Métodos



14. Se presenta la situación de una decisión en la que existen tres estados s_1, s_2 y s_3. Las probabilidades previas son $P(s_1) = 0.2$, $P(s_2) = 0.5$ y $P(s_3) = 0.3$. Con la información muestral $P(I \mid s_1) = 0.1$, $P(I \mid s_2) = 0.05$ y $P(I \mid s_3) = 0.2$. Calcule las probabilidades posteriores $P(s_1 \mid I)$, $P(s_2 \mid I)$ y $P(s_3 \mid I)$.
15. Las cantidades en una tabla de recompensa para un problema de decisión son ganancias. En este problema se tienen dos estados y tres alternativas de decisión. Las probabilidades previas de s_1 y s_2 son $P(s_1) = 0.8$, y $P(s_2) = 0.2$

	Estado	
Alternativa de decisión	s_1	s_2
d_1	15	10
d_2	10	12
d_3	8	20

a. ¿Cuál es la decisión óptima?
b. Hallar el VEIP.
c. Si se obtiene la información muestral y que $P(I \mid s_1) = 0.20$ y $P(I \mid s_2) = 0.75$. Halle las probabilidades posteriores $P(s_1 \mid I)$ y $P(s_2 \mid I)$. Con base en estas probabilidades recomiende una alternativa de decisión.

Aplicaciones

16. Para economizar, Rona y Jerry se han puesto de acuerdo para irse juntos, en el coche de uno de ellos, al trabajo. Rona prefiere irse por la avenida Queen, que aunque es un poco más larga, es más segura. Jeery prefiere ir por la autopista, porque es más rápido. Deciden que cuando la autopista esté muy congestionada se irán por la avenida Queen. En la siguiente tabla de recompensa se dan los tiempos en minutos de este recorrido

	Estados	
Alternativas de decisión	**Autopista sin tráfico** s_1	**Autopista congestionada** s_2
Avenida Queen, d_1	30	30
Autopista, d_2	25	45

De acuerdo con su experiencia, Rona y Jerry piensan que la probabilidad de que la autopista esté congestionada es 0.15.

Además, coinciden en que el clima parece afectar la circulación en la autopista Sea

$D =$ despejado
$N =$ nublado
$L =$ lluvioso

Las probabilidades condicionales son las siguientes

$$P(C \mid s_1) = 0.8 \quad P(O \mid s_1) = 0.2 \quad P(L \mid s_1) = 0.0$$
$$P(C \mid s_2) = 0.1 \quad P(O \mid s_2) = 0.3 \quad P(L \mid s_2) = 0.6$$

a. Use el teorema de Bayes para calcular la probabilidad de cada una de las condiciones y las probabilidades condicionales de que la autopista esté despejada, s_1, o congestionada, s_2, dadas cada una de las condiciones climáticas.
b. Presente el árbol de decisión para este problema.
c. Dé la estrategia óptima de decisión y el tiempo de viaje esperado.

17. La empresa Gorman Manufacturing Company tiene que decidir si fabrica una pieza en su planta de Milan, Michigan, o si la compra a un proveedor. Las ganancias dependerán de la demanda del producto. En la siguiente tabla de recompensa se presentan las ganancias proyectadas (en dólares).

	Estado		
Alternativa de decisión	**Demanda baja** s_1	**Demanda media** s_2	**Demanda alta** s_3
Fabricarla, d_1	−20	40	100
Comprarla, d_2	10	45	70

Las probabilidades de los estados son $P(s_1) = 0.35$, $P(s_2) = 0.35$ y $P(s_3) = 0.30$.

a. Use el árbol de decisión para recomendar una decisión.
b. Use el VEIP para determinar si Gorman deberá tratar de obtener una mejor estimación de la demanda.
c. Un estudio de mercado de la demanda potencial del producto dará como resultado que la situación es favorable (F) o desfavorable (D). Las probabilidades son:

$$P(F \mid s_1) = 0.10 \quad P(D \mid s_1) = 0.90$$
$$P(F \mid s_2) = 0.40 \quad P(D \mid s_2) = 0.60$$
$$P(F \mid s_3) = 0.60 \quad P(D \mid s_3) = 0.40$$

¿Cuál es la probabilidad de que la investigación de mercado arroje un informe favorable?
d. ¿Cuál es la estrategia óptima para Gorman?
e. ¿Cuál es el valor esperado de la información de la investigación de mercado?

Resumen

El análisis de decisión se usa para determinar una alternativa de decisión recomendada o una estrategia óptima de decisión ante un panorama de eventos futuros inciertos y riesgosos. El objetivo del análisis de decisión es identificar la mejor alternativa de decisión o la estrategia óptima de decisión dada cierta información acerca de los eventos inciertos y de las posibles consecuencias o recompensas. A los eventos futuros inciertos se les conoce como eventos aleatorios y a los resultados de los eventos aleatorios se les conoce como estados.

Se mostró el uso de las tablas de recompensa y de los árboles de decisión para estructurar un problema de decisión y para describir las relaciones entre las decisiones, los eventos aleatorios y las consecuencias. Con las estimaciones para las probabilidades de los estados, el método del valor esperado se empleó para identificar la alternativa de decisión recomendada o la estrategia de decisión.

En los casos en que se puede disponer de información muestral acerca de los eventos aleatorios, hay una secuencia de decisiones que tomar. Primero se decide si obtener, o no, la información muestral. Si la respuesta a esta decisión es sí, habrá que elaborar una estrategia óptima de decisión basada en la información muestral específica. En este caso, los árboles de decisión y el método del valor esperado pueden usarse para determinar la estrategia óptima de decisión.

El complemento de Excel TreePlan se encuentra en el disco compacto que se distribuye con el libro.

El complemento de Excel TreePlan suele emplearse para elaborar árboles de decisión y para resolver los problemas de decisión presentados en este capítulo. El software TreePlan y el manual para el uso de TreePlan se encuentran en el sitio de la Red de ASW. En el apéndice, al final del capítulo, se presenta un ejemplo en el que se muestra cómo usar TreePlan para resolver el problema de PDC presentado en la sección 21.1.

Glosario

Evento aleatorio Evento futuro incierto que afecta a la consecuencia, o recompensa, relacionada con una decisión.

Consecuencia Resultado obtenido de la elección de una alternativa de decisión y la ocurrencia de un evento aleatorio. A una medida de la consecuencia se le suele denominar recompensa.

Estados Los resultados posibles de los eventos aleatorios que afectan la recompensa correspondiente a una alternativa de decisión.

Recompensa Una medida de la consecuencia de una decisión, por ejemplo, ingresos, costos, tiempo. Para cada combinación de una alternativa de decisión y un estado está asociada una recompensa (consecuencia).

Tabla de recompensa Representación tabular de las recompensas en un problema de decisión.

Árbol de decisión Representación gráfica de un problema de decisión que muestra el carácter secuencial del proceso de toma de la decisión.

Nodo Punto de intersección o de encuentro en un diagrama de influencia o en un árbol de decisión.

Nodos de decisión Nodos que indican puntos en los que hay que tomar una decisión.

Nodos aleatorios Indican los puntos donde existe incertidumbre respecto a la ocurrencia de un evento.

Rama Líneas que indican las alternativas que salen de los nodos de decisión y resultados que salen de los nodos aleatorios.

Método del valor esperado Método para elegir una alternativa de decisión, que se basa en el valor esperado de cada alternativa de decisión. La alternativa de la decisión recomendada es la que proporciona el mejor valor esperado.

Valor esperado (VE) En un nodo aleatorio, es el promedio ponderado de las recompensas. Los pesos son las probabilidades de los estados.

Valor esperado de la información perfecta (VEIP) Valor esperado de la información que le diría, con exactitud, a quien tiene que tomar la decisión cuál es el estado que va a ocurrir (es decir, información perfecta).

Probabilidades previas Las probabilidades de los estados antes de obtener la información muestral.

Información muestral Información nueva obtenida a través de la investigación o de la experimentación y que permite actualizar o modificar las probabilidades de los estados.

Probabilidades posteriores Las probabilidades de los estados una vez modificadas las probabilidades previas con base en la información muestral.

Estrategia de decisión Estrategia en la que interviene una secuencia de decisiones y resultados aleatorios para obtener una decisión óptima a un problema de decisión.
Valor esperado de la información muestral (VEIM) Es la diferencia entre el valor esperado de una estrategia óptima basada en una información muestral y el "mejor" valor esperado sin ninguna información muestral.
Teorema de Bayes Teorema que permite el uso de la información muestral para modificar probabilidades previas.
Probabilidad condicional Probabilidad de un evento dado que se conoce el resultado de otro evento (posiblemente) relacionado.
Probabilidad conjunta Probabilidad de que tanto la información muestral como un determinado estado ocurran simultáneamente.

Fórmulas clave

Valor esperado

$$\text{VE}(d_i) = \sum_{j=1}^{N} P(s_j)V_{ij} \tag{21.3}$$

Valor esperado de la información perfecta

$$\text{VEIP} = |\text{VEcIP} - \text{VEsIP}| \tag{21.4}$$

Valor esperado de la información muestral

$$\text{VEIM} = |\text{VEcIM} - \text{VEsIM}| \tag{21.5}$$

Caso problema: Estrategia de defensa en un juicio

John Campbell, empleado de Manhattan Construction Company, afirma haberse lastimado la espalda como resultado de una caída que sufrió mientras reparaba el techo del edificio de departamentos Eastview. En una demanda contra Doug Reynolds, propietario del edificio Eastview, en la que solicita una indemnización por $1 500 000, John afirma que el techo tenía secciones podridas y que esa caída podía haberse evitado si el señor Reynolds hubiera informado del problema a Manhattan Construction. El señor Reynolds notificó de la demanda a su compañía de seguros, Allied Insurance. Allied tiene que defender al señor Reynolds y decidir las medidas que tomará respecto a la demanda.

Después de algunas declaraciones y discusiones entre ambas partes, John Campbell accedió a aceptar una indemnización de $750 000. De esta manera, una opción para Allied es pagarle a John los $750 000 y resolver el caso. Pero, Allied está tratando de hacerle a John una contraoferta por $400 000, esperando que acepte una cantidad menor evitándose así los costos y la pérdida de tiempo de un juicio. Las primeras investigaciones de Allied indican que el caso de John es un caso severo; a Allied le preocupa que rechace la contraoferta y prefiera irse a juicio. Los abogados de Allied analizan cuál puede ser la reacción de John si le hacen la contraoferta de $400 000.

Los abogados concluyen que puede haber tres reacciones de John frente a la contraoferta de los $400 000: 1) que John acepte la contraoferta y se cierre el caso; 2) que John rechace la contraoferta y prefiera que un tribunal determine el monto de la indemnización, o 3) que John haga a Allied una contraoferta por $600 000. En el caso que John haga una contraoferta, Allied ha de-

cidido que no hará más contraofertas; aceptarán la contraoferta de John por $600 000 o irán a los tribunales.

Si el caso llega a los tribunales, Allied prevé tres resultados posibles: 1) que el tribunal rechace la demanda de John y Allied no tenga que pagar nada; 2) que el tribunal esté a favor de John y le conceda una indemnización de $750 000, o 3) que el tribunal concluya que el caso de John es un caso severo y le conceda $1 500 000.

Consideraciones clave en el desarrollo de la estrategia de Allied, son las probabilidades correspondientes a las posibles respuestas de John a la contraoferta de Allied por los $400 000, así como las probabilidades de los tres posibles resultados en los tribunales. Los abogados de Allied consideran que la probabilidad de que John acepte la contraoferta por $400 000 de Allied es 0.10, la probabilidad de que John rechace la contraoferta de $400 000 es 0.40 y la probabilidad de que John haga una contraoferta a Allied por $600 000 es 0.50. Si el caso llega a los tribunales, los abogados consideran que la probabilidad de que el tribunal conceda a John una indemnización por $1 500 000 es 0.30, la probabilidad de que el tribunal conceda a John una indemnización por $750 000 es 0.50 y la probabilidad de que el tribunal no conceda a John ninguna indemnización es 0.20.

Informe administrativo

Realice un análisis del problema en que se encuentra Allied Insurance y elabore un informe en el que resuma sus hallazgos y recomendaciones. No deje de incluir en este informe lo siguiente:

1. Un árbol de decisión.
2. Una recomendación sobre si Allied debe aceptar la oferta inicial de John de resolver la demanda con $750 000.
3. La estrategia de decisión que deba seguir Allied si decide hacer una contraoferta a John por $400 000.
4. Un perfil de riesgos para la estrategia que recomienda.

Apéndice 21.1 Solución del problema PDC con TreePlan

TreePlan* es un complemento de Excel que se puede usar para elaborar árboles de decisión para problemas de análisis de decisión. El paquete de software se encuentra en el sitio de la Red de ASW, http://asw.swlwarning.com. También se encuentra un manual sobre cómo poner en marcha y usar TreePlan. En el ejemplo siguiente se muestra cómo usar TreePlan para construir un árbol de decisión para el problema de PDC, presentado en la sección 21.1. En la figura 21.11 se presenta el árbol de decisión para el problema de PDC.

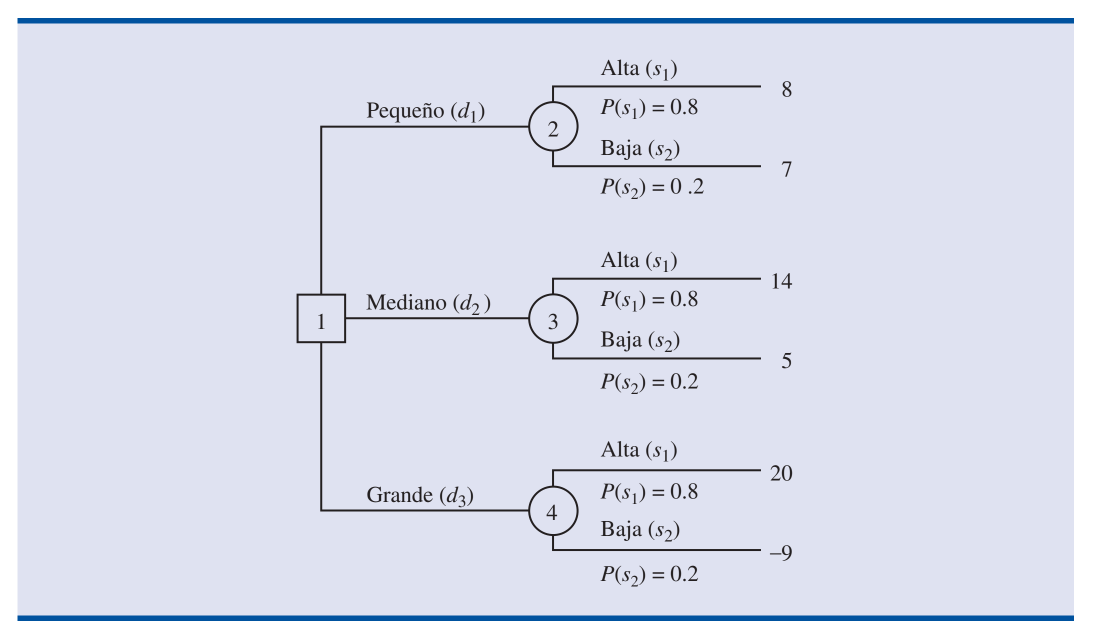
Para empezar: un primer árbol de decisión

Para empezar se supondrá que ya se ha instalado TreePlan y que ya se ha abierto un libro de Excel. Para obtener una versión en TreePlan del árbol de decisión para PDC se procede como sigue:

Paso 1. Seleccionar la celda A1
Paso 2. Seleccionar el menú **Herramientas** y elegir **Decision Tree**
Paso 3. Cuando aparezca el cuadro de diálogo TreePlan New:
Clic en **NewTree**

*TreePlan fue elaborado por el profesor Michael R. Middleton de la Universidad de San Francisco y modificado por el profesor James E. Smith de la Duke University. El sitio en la red de TreePlan es www.treeplan.com.

FIGURA 21.11 ÁRBOL DE DECISIÓN PARA PDC



Aparecerá un árbol de decisión con un nodo de decisión y dos ramas:

	A	B	C	D	E	F	G
1							
2				Decision 1			
3							0
4				0	0		
5		1					
6	0						
7				Decision 2			
8							0
9				0	0		

Agregar una rama

En el problema de PDC hay tres alternativas de decisión (un complejo pequeño, mediano o grande), de manera que es necesario agregarle al árbol una rama de decisión.

Paso 1. Seleccionar la celda B5
Paso 2. Seleccionar el menú **Herramientas** y elegir **Decision Tree**
Paso 3. Cuando aparezca el cuadro de diálogo TreePlan:
Seleccionar **Add branch**
Clic en **OK**

En la hoja de cálculo de Excel aparecerá un árbol modificado con tres ramas de decisión.

Dar nombre a las alternativas de decisión

A las alternativas de decisión se les puede dar un nombre al seleccionar las celdas que contienen los rótulos Decision 1, Decision 2 y Decision 3 e ingresando después los nombres correspondientes que se tienen en el problema de PDC, Grande (Large), Mediano (Medium) y Pequeño (Small). Una vez que se ha dado nombre a las alternativas, el árbol de decisión de PDC con tres ramas se verá como sigue:

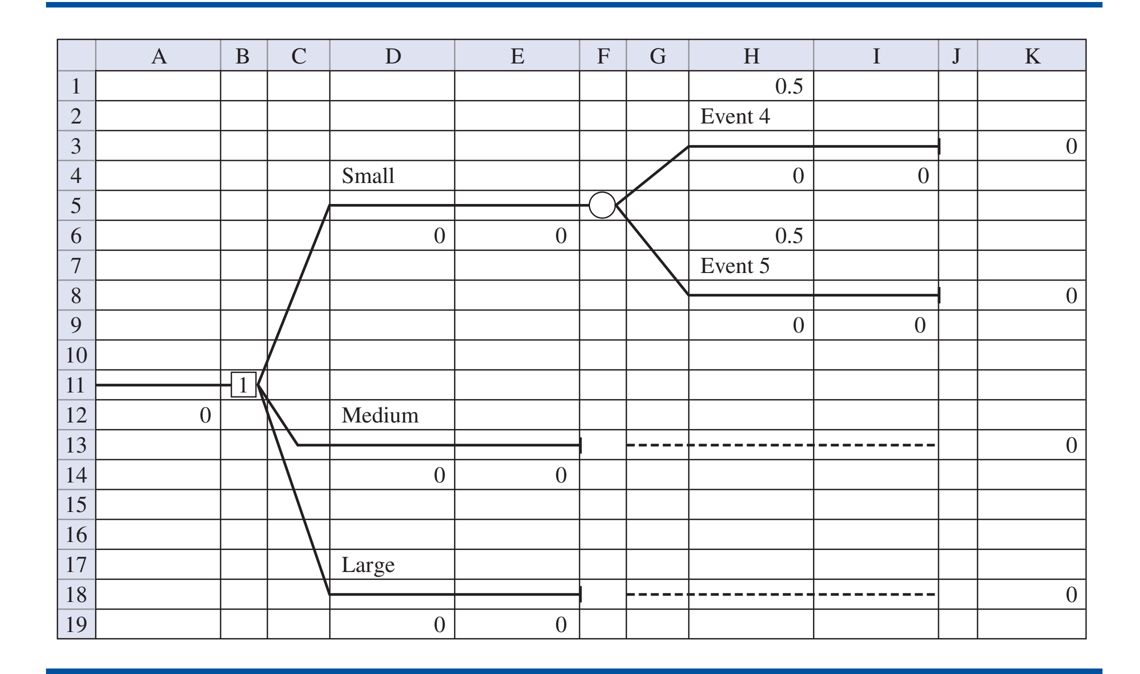
	A	B	C	D	E	F	G
1							
2				Small			
3							0
4				0	0		
5							
6							
7				Medium			
8		1					0
9	0			0	0		
10							
11							
12				Large			
13							0
14				0	0		

Agregar un nodo aleatorio

En el problema de PDC un evento aleatorio es la demanda de los condominios, la cual puede ser alta o baja. De manera que es necesario agregar un nodo aleatorio con dos ramas al final de cada rama de una alternativa de decisión.

Paso 1. Seleccionar la celda F3
Paso 2. Seleccionar el menú **Herramientas** y elegir **Decision Tree**
Paso 3. Cuando aparezca el cuadro de diálogo TreePlan New:
Seleccionar **Change to event node**
Seleccionar **Two** en la sección **Branches**
Clic en **OK**

Ahora aparecerá el árbol siguiente:



A continuación se seleccionan las celdas que contienen Event 4 y Event 5 y se les cambia el nombre a Alta (Strong) y Baja (Weak), para tener los nombres correspondientes a los estados del problema de PDC. Una vez hecho, se copia el árbol del nodo aleatorio en la celda F5 a las otras dos ramas de decisión para terminar la estructura del árbol de decisión para PDC.

Paso 1. Seleccionar la celda F5
Paso 2. Seleccionar el menú **Herramientas** y elegir **Decision Tree**
Paso 3. Cuando aparezca el cuadro de diálogo TreePlan:
Seleccionar **Copy subtree**
Clic en **OK**
Paso 4. Seleccionar la celda Fl5
Paso 5. Seleccionar el menú **Herramientas** y elegir **Decision Tree**
Paso 6. Cuando aparezca el cuadro de diálogo TreePlan:
Seleccionar **Paste subtree**
Clic en **OK**

Mediante este procedimiento de copiar/pegar (copy/paste) se coloca un nodo aleatorio al final de la rama de decisión Mediano (Medium). Repitiendo este procedimiento de copiar/pegar (copy/paste) con la rama de decisión Alta (Large) se obtiene la estructura completa del árbol de decisión que aparece en la figura 21.12.

FIGURA 21.12 ÁRBOL DE DECISIÓN DE PDC OBTENIDO MEDIANTE TREEPLAN

	A	B	C	D	E	F	G	H	I	J	K
1								0.5			
2								Strong			
3											0
4				Small				0	0		
5											
6				0	0			0.5			
7								Weak			
8											0
9								0	0		
10											
11								0.5			
12								Strong			
13											0
14				Medium				0	0		
15		1									
16	0			0	0			0.5			
17								Weak			
18											0
19								0	0		
20											
21								0.5			
22								Strong			
23											0
24				Large				0	0		
25											
26				0	0			0.5			
27								Weak			
28											0
29								0	0		

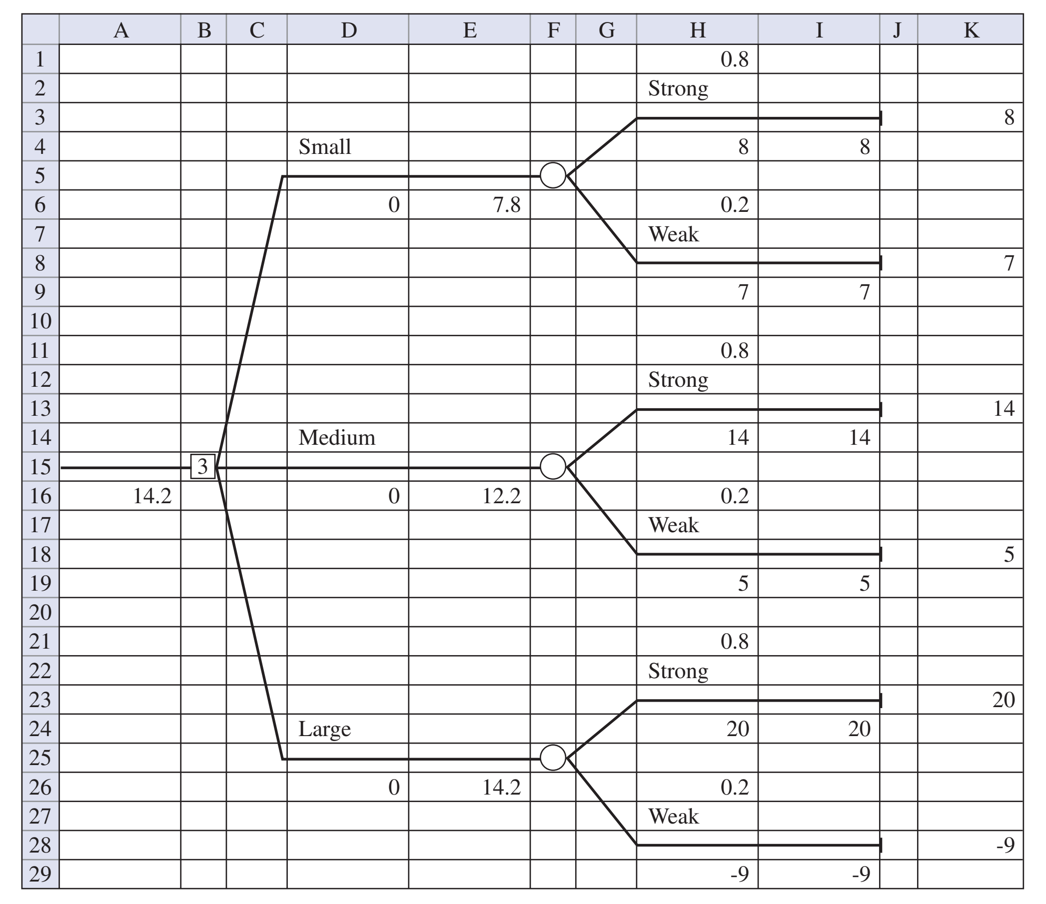
Inserción de probabilidades y recompensas

Con TreePlan puede insertar probabilidades y recompensas en el árbol de decisión. En la figura 21.12 se observa que TreePlan asigna, de manera automática, la probabilidad 0.5 a todos los estados. En el problema de PDC, la probabilidad de que la demanda sea alta es 0.8 y la probabilidad de que la demanda sea baja es 0.2. Si selecciona las celdas H1, H6, H11, H16, H21 y H26 se pueden insertar las probabilidades adecuadas. Las recompensas de los resultados aleatorios se han insertado en las celdas H4, H9, H14, H19, H24 y H29. Una vez insertadas las probabilidades y las recompensas, el árbol de decisión de PDC se verá como se muestra en la figura 21.13.

Observe que las recompensas aparecen también en el margen derecho del árbol de decisión. Las recompensas en el margen derecho se calculan mediante una fórmula que agrega las recompensas a todas las ramas que llevan al nodo terminal correspondiente. En el problema de PDC no hay recompensas para las ramas de las decisiones alternativas, por lo que en las celdas D6, D16 y D26 se ha dejado el valor cero que es el que aparece por defecto. Con esto queda terminado el árbol de decisión para PDC.

FIGURA 21.13 ÁRBOL DE DECISIÓN DE PDC CON RAMAS DE PROBABILIDAD Y RECOMPENSAS





Interpretación de los resultados

Si se insertan las probabilidades y las recompensas, TreePlan realiza automáticamente, en forma regresiva, los cálculos necesarios para obtener los valores esperados y determinar la solución óptima. Las decisiones óptimas se identifican mediante su número en el nodo de decisión correspondiente. En la figura 21.13, en el árbol de decisión de PDC el nodo de decisión se encuentra en la celda B15. Observe que en ese nodo aparece un 3, lo que indica que la rama 3 de las alternativas de decisión proporciona la decisión óptima. Es decir, el análisis de decisión recomienda que PDC construya el complejo grande. El valor esperado de esta decisión aparece al principio del árbol en la celda A16. Como se ve, el valor óptimo esperado es $14.2 millones. El valor esperado de las otras alternativas aparece al final de las correspondientes ramas de decisión. En las celdas E6 y E16 se ve que el valor esperado para el complejo pequeño es $7.8 millones y el valor esperado para el complejo mediano es $12.2 millones.

Otras opciones

Por defecto, el objetivo de TreePlan es una maximización. Cuando el objetivo es una minimización se siguen los pasos que se presentan a continuación:

Paso 1. Seleccionar el menú **Herramientas** y elegir **Decision Tree**
Paso 2. Seleccionar **Options**
Paso 3. Elegir **Minimize (costs)**
Clic en **OK**

Con TreePlan se pueden modificar las probabilidades y las recompensas y ver rápidamente el impacto de estos cambios sobre la solución óptima. Con el tipo “y si” de análisis de sensibilidad se pueden identificar modificaciones de las probabilidades y de las recompensas que modifiquen la decisión óptima. Además, como TreePlan es un complemento de Excel, la mayor parte de las posibilidades de Excel pueden emplearse. Por ejemplo, se pueden usar negritas para resaltar el nombre de la alternativa óptima de decisión. En el manual de TreePlan se encuentran otras muchas opciones con las que cuenta TreePlan. Paquetes de software como TreePlan facilitan el análisis meticuloso de los problemas de decisión.

CAPÍTULO 22



Encuestas muestrales

CONTENIDO



LA ESTADÍSTICA *en* LA PRÁCTICA

DUKE ENERGY*
CHARLOTTE, CAROLINA DEL NORTE

Duke Energy es una empresa diversificada de energía con un portafolio de negocios en gas natural y electricidad y una empresa inmobiliaria afiliada. En el 2006, Duke Energy se fusionó con Cinergy de Cincinnati, Ohio, y formaron una de las empresas de energía más grandes de Estados Unidos, con un activo que asciende a más de $70 mil millones. En la actualidad, Duke Energy da servicio a más de 5.5 millones de usuarios de gas y electricidad en Carolina del Norte, Carolina del Sur, Ohio, Kentucky, Indiana y Ontario, Canadá.

Para dar un mejor servicio, Duke Energy constantemente está atenta a las necesidades emergentes de sus clientes. En el ejemplo siguiente se verá cómo esta empresa realizó una encuesta acerca de las características de los edificios para conocer mejor los requerimientos de energía de los edificios comerciales en el área de servicio de Cincinnati, Ohio.

La empresa buscó información diversa acerca de los edificios comerciales como arquitectura, cantidad de empleados, uso final dado a la energía, antigüedad del edificio, tipo de materiales de construcción y medidas para la conservación de la energía. Durante los preparativos de la encuesta, los analistas determinaron que en el área de servicio de Cincinnati había aproximadamente 27 000 edificios comerciales en función. De acuerdo con los recursos disponibles y con la precisión deseada para la encuesta, se recomendó tomar una muestra de 616 edificios comerciales.

El tipo de muestreo que se eligió fue un muestreo aleatorio simple estratificado. La empresa contaba con registros sobre el consumo de energía total en los años recientes de cada uno de los edificios en el área de servicio de Cincinnati y, dado que muchas de las características que interesaban de los edificios (tamaño, cantidad de empleados, etc.) estaban relacionadas con el consumo, ése fue el criterio empleado para dividir la población de edificios en seis estratos.

El primer estrato estaba constituido por los edificios que eran los 100 principales consumidores de energía; todos los edificios de este estrato fueron incluidos en la muestra. Aunque estos edificios constituían únicamente el 0.2% de la población, consumían el 14.4% de toda la energía eléctrica. De los otros estrados, el número de edificios muestreados se determinó en función de la obtención de la mayor precisión posible por costo unitario.

La empresa elaboró un cuestionario que se probó antes de realizar la encuesta. Los datos se obtuvieron a través de entrevistas personales. De los 616 edificios comerciales de la muestra se obtuvieron 526 cuestionarios completamente contestados. Esta tasa de respuesta de 85.4% fue excelente. La empresa usó los resultados de la encuesta para pronosticar la demanda de energía y para mejorar el servicio prestado a sus clientes comerciales.

En este capítulo, el lector conocerá los tópicos que consideran los estadísticos para el diseño y realización de una encuesta muestral como la realizada por Duke Energy. Las encuestas muestrales suelen emplearse para obtener perfiles de los clientes de una empresa; también son empleados por los gobiernos y por otras instituciones para conocer diversos segmentos de la población.



En Cincinnati, Ohio, se llevó a cabo una encuesta muestral sobre las necesidades de electricidad en los edificios comerciales. © Getty Images/PhotoDisc.



* Los autores agradecen a Jim Ruddle de Duke Energy por proporcionar este artículo para *La Estadística en la práctica*.

22.1 Terminología empleada en las encuestas muestrales

En el capítulo 1 se dieron las siguientes definiciones de elemento, población y muestra.

- **Elemento** es la entidad de la que se toman los datos.
- **Población** es la colección de todos los elementos que interesan.
- **Muestra** es un subconjunto de la población.

Para ilustrar estos conceptos considere la situación siguiente. Dunning Microsystems, Inc. (DMI), fabricante de computadoras personales y periféricos, desea obtener datos acerca de las características de las personas que le han comprado sus computadoras personales. Para esto, debe realizar una encuesta muestral a los poseedores de una computadora personal DMI. En esta encuesta muestral los *elementos* son cada uno de los individuos que hayan comprado una computadora personal DMI. La *población* es el conjunto de todas las personas que hayan comprado una computadora personal DMI y la *muestra* será el subconjunto de poseedores de una computadora personal que se tome para la encuesta.

En las encuestas muestrales es necesario distinguir entre la población objetivo y la población muestreada. La **población objetivo** es la población acerca de la cual se desean hacer inferencias, mientras que la **población muestreada** es la población de la que, realmente, se toma la muestra. Es importante entender que estas dos poblaciones no siempre son una misma. En el ejemplo de DMI, la población objetivo consta de todas las personas que han comprado una computadora personal DMI. La población muestreada, en cambio, puede que sea, por ejemplo, todos los poseedores de una computadora personal DMI que hayan enviado a DMI la tarjeta de registro para la garantía. No todos los que compran una computadora personal DMI envían la tarjeta de registro para la garantía, de manera que la población muestreada es diferente de la población objetivo.

Las inferencias obtenidas a partir de una muestra son válidas, si la población muestreada es representativa de la población objetivo.

Las conclusiones que se obtienen de una encuesta muestral sólo son válidas para la población muestral. El que estas conclusiones puedan o no ampliarse a la población objetivo depende del criterio del analista. El punto clave es si entre la población muestreada y la población objetivo existe una semejanza suficiente respecto a la característica de interés como para permitir ampliar las conclusiones.

Antes del muestreo, se divide la población en **unidades muestrales**. En algunos casos las unidades muestrales son simplemente los elementos. En otros casos, las unidades muestrales son grupos de elementos. Por ejemplo, suponga que se desea hacer una encuesta a los ingenieros que trabajan en el diseño de sistemas de calefacción y de aire acondicionado para edificios comerciales. Si se tuviera una lista de todos estos ingenieros, las unidades muestrales serían los ingenieros que se desea investigar. Si no se cuenta con tal lista, es necesario hallar otra alternativa. Una alternativa puede ser la lista de las empresas de ingeniería que se dedican al diseño de sistemas de calefacción y aire acondicionado que se encuentran en un directorio telefónico comercial. Dada tal lista se toma una muestra de estas empresas para la encuesta; en cada empresa tomada para la encuesta se entrevista a todos los ingenieros. En este caso las unidades muestrales serán las empresas de ingeniería y los elementos serán los ingenieros entrevistados.

A la lista de las unidades muestrales tomadas para un estudio particular se le conoce como el **marco**. En la encuesta a los ingenieros el marco está definido por todas las empresas de ingeniería enumeradas en el directorio telefónico; el marco no es una lista de todos los ingenieros porque no se cuenta con tal lista. El marco que se elija y, por tanto, la definición de las unidades muestrales, suele estar determinado por la lista de que se disponga y la confiabilidad de la misma. En la práctica la elección del marco suele ser uno de los pasos más difíciles e importantes al realizar una encuesta muestral.

22.2 Tipos de encuestas y métodos de muestreo

Los tres tipos de encuestas muestrales más comunes son las encuestas por correo, las encuestas por teléfono y las encuestas a través de entrevistas personales. Hay otros tipos de investigaciones que se emplean para recabar datos en los que no se emplean cuestionarios. Por ejemplo, para muestrear el inventario de bienes de una empresa con objeto de estimar el valor de inventario en el balance general de la empresa suele contratarse a una empresa de contadores. En estas investigaciones, una persona simplemente cuenta los artículos y anota los resultados.

En las encuestas en que se usan cuestionarios, el diseño del cuestionario es relevante. Al hacer el diseño de un cuestionario hay que resistirse a incluir preguntas que *pueden* ser de interés, ya que cada pregunta agregada al cuestionario lo hace más largo. Los cuestionarios largos no sólo conducen al cansancio del entrevistado, sino también al del entrevistador, en especial cuando se trata de encuestas por correo o por teléfono. Cuando se emplean entrevistas personales, es po-

Los costos de las encuestas por correo o por teléfono son más bajos, pero las entrevistas personales, cuando se cuenta con entrevistadores bien capacitados, suelen dar tasas de respuesta más altas y permitir cuestionarios más largos. En el caso del estudio presentado en el artículo de La estadística en la práctica de este capítulo, dada la cantidad de datos que se deseaba recabar de cada elemento, la única posibilidad de hacerlo era mediante entrevistas personales.

sible hacer cuestionarios más largos y más complejos. Ya existe una gran cantidad de conocimientos sobre la redacción y secuencia de las preguntas para un cuestionario, así como sobre la manera de agruparlas. Estos temas corresponden a libros especializados sobre encuestas muestrales; en la bibliografía se citan varias fuentes en las que se encontrará este tipo de información.

Las encuestas muestrales también se clasifican de acuerdo con el método de muestreo que se utilice. Los **muestreos probabilísticos** permiten calcular la probabilidad de obtener cada una de las posibles muestras; en los **muestreos no probabilísticos** esto no es posible. Los métodos no probabilísticos de muestreo no deben usarse cuando el investigador desea determinar la precisión de las estimaciones. En cambio, los métodos probabilísticos de muestreo se emplean para obtener intervalos de confianza con los que se pueden obtener límites para el error muestral. En las secciones siguientes se estudiarán cuatro de los métodos de muestreo probabilístico más usados: el muestreo aleatorio simple, el muestreo aleatorio simple estratificado, el muestreo por conglomerados y el muestreo sistemático.

Aunque los especialistas en estadística prefieren usar los métodos probabilísticos de muestreo, los métodos no probabilísticos de muestreo también suelen ser necesarios. Las ventajas de los métodos de muestreo no probabilísticos son su bajo costo y su fácil realización. La desventaja es que no se puede decir de una manera estadística válida cuál es la precisión de la estimación. Dos de los métodos probabilísticos más usados son el muestreo de conveniencia y el muestreo subjetivo.

En el **muestreo de conveniencia** las unidades que se toman en la muestra, se toman por su accesibilidad. Por ejemplo, cuando un profesor de una universidad que realiza una investigación suele solicitar alumnos voluntarios que participen en el estudio, estos alumnos participan en la muestra sólo porque son alumnos del profesor. En este caso a la muestra de estudiantes se le conoce como muestra de conveniencia. En algunos casos el muestreo de conveniencia es la única posibilidad práctica. Por ejemplo, para muestrear un cargamento de naranjas, el investigador tomará de manera aleatoria naranjas de varias cajas ya que no sería práctico etiquetar todas las naranjas del cargamento para obtener un marco y emplear un método de muestreo probabilístico. Otros ejemplos de muestreos de conveniencia son los estudios sobre la flora y fauna salvajes y los paneles de voluntarios en las investigaciones de mercado.

Aun cuando el muestreo de conveniencia es una manera relativamente sencilla de seleccionar una muestra y obtener los datos deseados, es imposible estimar la "bondad" de los estadísticos muestrales obtenidos como estimaciones de los parámetros poblacionales que interesan. Un muestreo de conveniencia puede o no dar buenos resultados; no hay ningún procedimiento estadísticamente justificado que permita hacer inferencias muestrales a partir de los resultados muestrales. A pesar de esto, algunas veces, los investigadores aplican métodos probabilísticos, diseñados para muestras probabilísticas, a datos obtenidos mediante un muestreo de conveniencia. En tales casos, el investigador suele argumentar que la muestra de conveniencia puede ser considerada como una muestra aleatoria en el sentido de que es representativa de la población. Pero, hay que cuestionar este argumento; se debe ser muy cuidadoso al emplear muestras de conveniencia para hacer inferencias estadísticas acerca de los parámetros poblacionales.

En el caso de los métodos de muestreo no probabilístico, cuando se pueden emplear métodos que garanticen que se ha obtenido una muestra representativa, las estimaciones puntuales basadas en la muestra pueden ser útiles. Sin embargo, incluso en estos casos no se puede conocer la precisión de los resultados.

En la técnica de muestreo conocida como **muestreo subjetivo**, una persona, con conocimientos en la materia de estudio, selecciona las unidades muestrales que considera más representativas de la población. El muestreo subjetivo suele ser una manera relativamente sencilla para tomar una muestra, sin embargo, los usuarios de los resultados de tales encuestas deben aceptar que la calidad de los resultados es dependiente del criterio de la persona que selecciona la muestra. Por tanto, se debe tener mucho cuidado al usar muestras subjetivas para hacer inferencias estadísticas acerca de los parámetros poblacionales. En general, conviene no hacer aseveraciones estadísticas acerca de la precisión de los resultados obtenidos de una muestra subjetiva.

Para tomar una muestra se pueden usar tanto métodos probabilísticos como no probabilísticos. La ventaja de los métodos no probabilísticos es que, por lo general, no son caros y son fáciles de usar. Pero cuando se necesita indicar la precisión de las estimaciones, será necesario emplear métodos probabilísticos de muestreo. En casi todas las encuestas muestrales grandes se emplean métodos probabilísticos.



Errores en una encuesta

Al realizar una encuesta se pueden presentar dos tipos de errores. Uno, el error muestral, que es la magnitud de la diferencia entre el estimador puntual insesgado obtenido de la muestra y el parámetro poblacional. En otras palabras, el error muestral es el error que se presenta debido a que no se investigan todos los elementos de la población. El segundo error es el error no muestral, que se refiere a todos los demás tipos de errores que se presentan cuando se realiza una encuesta, como errores de medición, errores del entrevistador y errores de procesamiento. Los errores muestrales sólo pueden presentarse en una encuesta muestral; los errores no muestrales ocurren tanto en un censo como en una encuesta muestral.

Errores no muestrales

Uno de los errores no muestrales más comunes se presenta cuando una característica de interés es medida de forma incorrecta. Los errores de medición ocurren tanto en un censo como en una encuesta muestral. En cualquier tipo de encuesta es necesario tener cuidado de que todos los instrumentos de medición (por ejemplo, los cuestionarios) estén adecuadamente calibrados y de que las personas que hagan las mediciones estén debidamente capacitadas. Poner atención a los detalles es la mejor precaución en la mayor parte de las situaciones.

Los errores no muestrales se minimizan mediante una capacitación adecuada de los entrevistadores, un buen diseño de los cuestionarios, que deben ser probados antes de ser empleados en la encuesta, y cuidado en el proceso de codificación y transferencia de los datos a la computadora.

Los errores debidos a la falta de respuestas preocupan tanto al especialista en estadística, que es el responsable del diseño de la encuesta, como al ejecutivo que usará los resultados de la misma. Este tipo de error no muestral se presenta siempre que no es posible obtener, de algunas de las unidades de la encuesta, los datos deseados, o cuando únicamente se obtienen datos parciales. Un problema más serio es cuando se crea un sesgo. Por ejemplo, si se realizan entrevistas para evaluar la opinión de las mujeres respecto de que las mujeres trabajen fuera de casa y se llama a los hogares únicamente durante el día, se creará un sesgo obvio, debido a que las mujeres que trabajan fuera de casa quedarán excluidas de la muestra.

En el censo llevado a cabo en Estados Unidos en 1990, 25.9% de los hogares no respondieron. En el censo de 2000 se hizo un estudio muestral de los que no respondían al censo con objeto de estimar las características de esta porción de la población.

En encuestas técnicas son comunes los errores no muestrales que se deben a falta de conocimientos de los entrevistados. Por ejemplo, suponga que se hace una encuesta entre los administradores de edificios para obtener información detallada acerca del tipo de sistemas de ventilación que se usan en los edificios de oficinas. Los administradores de edificios grandes de oficinas tendrán buenos conocimientos acerca de tales sistemas, ya que es probable que hayan asistido a seminarios y obtengan apoyos para mantenerse informados y al día. En cambio, es posible que los administradores de edificios pequeños tengan menos conocimientos acerca de tales sistemas, debido a la gran variedad de tareas que deben realizar. Esta diferencia en los conocimientos afecta significativamente los resultados de la encuesta.

Otros dos tipos de errores no muestrales son el error de selección y el error de procesamiento. El error de selección se presenta cuando en la muestra se incluye algún elemento que no sea adecuado. Suponga que se diseña una encuesta muestral para obtener el perfil de un hombre con barba; si algunos entrevistadores entienden que entre los "hombres con barba" están comprendidos los hombres con bigote, mientras que otros entienden que no lo están, los datos resultantes serán deficientes. Los errores de procesamiento se presentan cuando los datos son anotados con incorrecciones o cuando son transferidos de manera incorrecta, por ejemplo, de los cuestionarios a la computadora.

Aun cuando algunos de los errores no muestrales se presentan en la mayor parte de las encuestas, es posible minimizarlos mediante una planeación cuidadosa. Debe tener cuidado de que haya una estrecha correspondencia entre la población muestreada y la población objetivo; que se sigan los buenos principios para la formulación de cuestionarios; que los entrevistadores estén bien capacitados, etc. En el informe final de una encuesta es recomendable incluir un análisis sobre el impacto que pueden tener los errores no muestrales sobre los resultados.

Error muestral

Recuerde la encuesta muestral de Dunning Microsystems (DMI). Suponga que DMI desea estimar la edad promedio de las personas que compran una computadora personal. Si se pudiera investigar a toda la población de personas que poseen una computadora DMI (hacer un censo) y

no se cometiera ningún error no muestral, se podría determinar esta edad promedio con toda exactitud. Pero, ¿qué pasa si no se puede investigar el 100% de todos los propietarios de una computadora DMI? En este caso, es posible que exista alguna diferencia entre la media muestral y la media poblacional; al valor absoluto de esta diferencia se le conoce como error muestral. En la práctica no es posible determinar cuál es el error muestral en una muestra determinada, ya que no es posible conocer la media poblacional, sin embargo, sí es posible dar una estimación probabilística acerca del tamaño del error muestral.

El error muestral se minimiza al elegir un diseño adecuado para la muestra.

Como ya se dijo, el error muestral se debe a que la encuesta se hace a partir de una muestra y no de toda la población. Aun cuando el error muestral no puede evitarse, sí es controlable. Una manera de controlar este tipo de error es elegir un método o diseño apropiado de muestreo. En las secciones siguientes se verán cuatro métodos de muestreo probabilístico: aleatorio simple, aleatorio estratificado, por conglomerados y sistemático.



22.4 Muestreo aleatorio simple

La definición de muestreo aleatorio simple se presentó en el capítulo 7:

> Una muestra aleatoria simple de tamaño n tomada de una población finita de tamaño N es una muestra que se elige de tal manera que todas las muestras posibles de tamaño n tengan la misma probabilidad de ser elegidas.

Para realizar una encuesta muestral usando el **muestreo aleatorio simple**, se empieza por elaborar un marco o lista de todos los elementos de la población muestral. A continuación se emplea un procedimiento de selección que se basa en el uso de números aleatorios, para garantizar que todos los elementos de la población muestral tengan la misma probabilidad de ser elegidos para la muestra. En esta sección se verá cómo se obtienen estimaciones de la media, del total y de la proporción poblacionales cuando en una encuesta muestral se usa el muestreo aleatorio simple.

Media poblacional

En el capítulo 8 se vio que la media muestral $\bar{x}$ es una estimación de la media poblacional μ y que la desviación estándar muestral s es una estimación de la desviación estándar poblacional σ. Un intervalo de estimación para μ, dada una muestra de tamaño n y empleando la distribución t, es el siguiente.

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{22.1}$$

En la expresión (22.1), $s/\sqrt{n}$ es la estimación de $\sigma_{\bar{x}}$, el error estándar de la media.

Cuando la muestra aleatoria simple de tamaño n se toma de una población finita de tamaño N, la estimación del error estándar de la media se obtiene mediante la fórmula siguiente

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right) \tag{22.2}$$

Al usar $s_{\bar{x}}$ como estimación de $\sigma_{\bar{x}}$ el intervalo de estimación para la media poblacional se convierte en

$$\bar{x} \pm t_{\alpha/2}s_{\bar{x}} \tag{22.3}$$

En las encuestas muestrales se acostumbra emplear el valor $t = 2$ para obtener una estimación por intervalo. Por tanto, cuando se emplea el muestreo aleatorio simple, el intervalo de confianza de aproximadamente 95% para estimar la media poblacional es el dado por la expresión siguiente.

INTERVALO DE CONFIANZA DE APROXIMADAMENTE 95% PARA ESTIMAR LA MEDIA POBLACIONAL

$$\bar{x} \pm 2s_{\bar{x}} \tag{22.4}$$

Considere, por ejemplo, el caso de la editorial de la revista *Great Lakes Recreation,* una revista regional especializada en navegación y pesca. En la actualidad, la revista cuenta con 8 000 suscriptores. En una muestra aleatoria simple de $n = 484$ suscriptores el ingreso anual medio encontrado fue $30 500 y la desviación estándar $7 040. Una estimación insesgada del ingreso anual medio de todos los suscriptores es $\bar{x} = \$30\,500$. Con estos resultados muestrales y la ecuación (22.2) se obtiene la estimación siguiente para el error estándar de la media.

$$s_{\bar{x}} = \sqrt{\frac{8000 - 484}{8000}}\left(\frac{7040}{\sqrt{484}}\right) = 310$$

Por tanto, de acuerdo con la fórmula (22.4), un intervalo de confianza de aproximadamente 95% para el ingreso anual medio de los suscriptores, es

$$30\,500 \pm 2(310) = 30\,500 \pm 620$$

es decir, $29 880 a $31 120.

El número que se le suma y se le resta a la estimación puntual para obtener el intervalo de estimación, se conoce como **cota del error muestral**. Por ejemplo, en la encuesta muestral de *Great Lakes Recreation,* una estimación del error estándar del estimador puntual es $s_{\bar{x}} = \$310$, y la cota del error muestral es $2(\$310) = \$620$.

El procedimiento anterior también sirve para calcular intervalos de estimación para otros parámetros poblacionales, como, por ejemplo, para el total poblacional y para la proporción poblacional. En estos casos, un intervalo de confianza de aproximadamente 95% puede expresarse de la manera siguiente

Estimador puntual ± 2(Estimación del error estándar del estimador puntual)

Total poblacional

Considere el problema que se le plantea a la empresa Northeast Electric and Gas (NEG). Como parte de un estudio sobre el consumo de energía, NEG necesita estimar el área *total*, en pies cuadrados, de las 500 escuelas públicas en su área de servicio. Esta área total de las 500 escuelas públicas se denotará como X; en otras palabras, X denota la población total. Observe que si se conociera μ, el promedio en pies cuadrados de las 500 escuelas públicas, al multiplicar N por μ se obtendría el valor de X. Pero, como no se conoce μ, una estimación puntual de X es la que se obtiene al multiplicar N por $\bar{x}$. El estimador puntual de X se denota $\hat{X}$.

ESTIMADOR PUNTUAL DEL TOTAL POBLACIONAL

$$\hat{X} = N\bar{x} \tag{22.5}$$

La estimación del error estándar de este estimador puntual está dada por

$$s_{\hat{X}} = Ns_{\bar{x}} \tag{22.6}$$

donde

$$s_{\bar{x}} = \sqrt{\frac{N - n}{N}}\left(\frac{s}{\sqrt{n}}\right) \tag{22.7}$$

Observe que la ecuación (22.7) es la fórmula obtenida para la estimación de error estándar de la media. Para obtener un intervalo de confianza de aproximadamente 95% para estimar el total poblacional se emplea este error estándar y la ecuación (22.6).

INTERVALO DE CONFIANZA DE APROXIMADAMENTE 95% PARA ESTIMAR EL TOTAL POBLACIONAL

$$N\bar{x} \pm 2s_{\hat{X}} \quad \textbf{(22.8)}$$

Suponga que en el estudio de NEG se toma una muestra aleatoria de $n = 50$ escuelas públicas de la población de $N = 500$ escuelas; la media muestral es $\bar{x} = 22\,000$ pies cuadrados y la desviación estándar muestral es $s = 4\,000$ pies cuadrados. Mediante la ecuación (22.5) se obtiene

$$\hat{X} = (500)(22\,000) = 11\,000\,000$$

Para obtener una estimación del error estándar de la media se emplea la ecuación (22.7).

$$s_{\bar{x}} = \sqrt{\frac{500 - 50}{500}}\left(\frac{4\,000}{\sqrt{50}}\right) = 536.66$$

Después, con la ecuación (22.6), se obtiene una estimación del error estándar de $\hat{X}$.

$$s_{\hat{X}} = (500)(536.66) = 268\,330$$

Por tanto, con la expresión (22.8), se encuentra que un intervalo de confianza de aproximadamente 95% para el total de pies cuadrados de las 500 escuelas públicas en el área de servicio de NEG es

$$11\,000\,000 \pm 2(268\,330) = 11\,000\,000 \pm 536\,660$$

es decir, 10 463 340 a 11 536 660 pies cuadrados.

Proporción poblacional

Una proporción poblacional p es la fracción de elementos de la población que posee alguna característica de interés. Por ejemplo, en un estudio de investigación de mercado el interés puede ser la proporción de consumidores que prefieren determinada marca de un producto. La proporción muestral $\bar{p}$ es un estimador puntual insesgado de la proporción poblacional. Un estimador del error estándar de la proporción es el dado por

$$s_{\bar{p}} = \sqrt{\left(\frac{N - n}{N}\right)\left(\frac{\bar{p}(1 - \bar{p})}{n - 1}\right)} \quad \textbf{(22.9)}$$

Un intervalo de confianza de aproximadamente 95% para estimar la proporción muestral es el dado por la expresión siguiente.

INTERVALO DE CONFIANZA DE APROXIMADAMENTE 95% PARA ESTIMAR LA PROPORCIÓN POBLACIONAL

$$\bar{p} \pm 2s_{\bar{p}} \quad \textbf{(22.10)}$$

Como ejemplo, suponga que en el problema de muestreo de Northeast Electric and Gas, también quiere estimar la proporción de las 500 escuelas públicas, en su área de servicio, que emplean gas natural para la calefacción. Si 35 de las 50 escuelas muestreadas indican que usan gas natural, la estimación puntual de la proporción poblacional, en las 500 escuelas de la población que usan gas natural, es $\bar{p} = 35/50 = 0.70$. Mediante la ecuación (22.9), se calcula la estimación del error estándar de la proporción.

$$s_{\bar{p}} = \sqrt{\left(\frac{500 - 50}{500}\right)\left(\frac{0.7(1-0.7)}{50 - 1}\right)} = 0.0621$$

Por tanto, con la expresión (22.10) se encuentra que un intervalo de confianza de aproximadamente 95% para la proporción poblacional es

$$0.7 \pm 2(0.0621) = 0.7 \pm 0.1242$$

es decir, 0.5758 a 0.8242.

Como se ve en este ejemplo, en una estimación de la proporción poblacional la amplitud del intervalo de confianza puede ser bastante grande. En general, para obtener estimaciones precisas de las proporciones poblacionales se necesitan tamaños de muestra grandes. Louis Harris & Associates en un informe de una encuesta muestral de 529 inversionistas de fondos mutualistas, dice: “Los resultados deben tener una exactitud de 4.3 puntos porcentuales.” Esto significa que el intervalo de confianza de aproximadamente 95% tiene una amplitud de 0.086. En poblaciones grandes, muestras de $n = 200$, o más, son frecuentes.

Determinación del tamaño de la muestra

Una consideración importante en el diseño de la muestra es la elección de su tamaño. Lo mejor suele ser un compromiso entre costo y precisión. Las muestras mayores permiten una mejor precisión (cotas más estrechas del error muestral) pero son más costosas. Con frecuencia, lo que dicta el tamaño de la muestra es el presupuesto con que se cuenta para el proyecto. En otras ocasiones, el tamaño de la muestra debe ser lo suficientemente grande para que permita obtener un determinado nivel de precisión.

El método que suele emplearse para determinar el tamaño de la muestra es, primero, especificar la precisión deseada y después determinar el menor tamaño de muestra con el que se obtiene esa precisión. En el presente contexto, el término *precisión* se refiere al tamaño del intervalo de confianza aproximado; intervalos de confianza más pequeños proporcionan mayor precisión. Como el tamaño del intervalo de confianza aproximado depende de la cota B del error muestral, elegir un nivel de precisión equivale a elegir un valor para B. A continuación se muestra este método para la elección del tamaño de muestra necesario para estimar la media poblacional.

La ecuación (22.2) indica que la estimación del error estándar de la media es

$$s_{\bar{x}} = \sqrt{\frac{N - n}{N}}\left(\frac{s}{\sqrt{n}}\right)$$

Recuerde que la cota del error muestral es “2 multiplicado por la estimación del error estándar del estimador puntual”. Por tanto,

$$B = 2\sqrt{\frac{N - n}{N}}\left(\frac{s}{\sqrt{n}}\right) \tag{22.11}$$

Al despejar n en la ecuación (22.11), se obtiene una cota del error muestral igual a B. De esta manera

$$n = \frac{Ns^2}{N\left(\frac{B^2}{4}\right) + s^2} \tag{22.12}$$

Una vez elegido el nivel de precisión (un valor para B), al aplicar la ecuación (22.12) se obtiene el valor de n que permite obtener la precisión deseada. Pero el empleo de la ecuación (22.12) para elegir el valor de n presenta algunos problemas, pues además de especificar el valor de la cota deseada B del error muestral, se necesita el valor de la varianza muestral s^2, pero s^2 no se puede conocer sino hasta que se tome la muestra.

Cochran,* sugiere varias maneras prácticas para obtener el valor de s^2. Tres de ellas son las siguientes:

1. Tomar la muestra en dos etapas. Usar en la ecuación (22.12) el valor de s^2 hallado en la etapa 1; el valor que se obtenga para n es el tamaño que debe tener la muestra. Después, en la etapa 2 tomar el número de unidades adicionales necesarias para alcanzar el tamaño total de la muestra, determinada en la etapa 1.
2. Usar los resultados de una encuesta piloto o de una prueba preliminar para estimar s^2.
3. Usar la información de una muestra previa.

A continuación se verá un ejemplo en el que se estiman los salarios medios iniciales de los egresados de una determinada universidad. Suponga que hay 5 000 egresados y se quiere obtener un intervalo de confianza de aproximadamente 95% cuya amplitud sea, a lo más, $1 000. Para obtener este intervalo de confianza se requiere que $B = 500$. Antes de usar la ecuación (22.12) para estimar el tamaño de la muestra, se necesita estimar s^2. Suponga que en un estudio similar realizado el año anterior se encontró $s = \$3\,000$. Para estimar s^2 se pueden emplear los datos de esa muestra anterior. Ahora ya se puede usar la ecuación (22.12), con $B = 500$, $s = 3\,000$ y $N = 5\,000$, para determinar el tamaño de la muestra.

$$n = \frac{5\,000(3\,000)^2}{5\,000\left(\frac{(500)^2}{4}\right) + (3\,000)^2}$$
$$= 139.97$$

Al redondear hacia arriba, se halla que para obtener un intervalo de confianza de aproximadamente 95% cuya amplitud sea $1 000, se necesita que el tamaño de la muestra sea de 140. Pero hay que tener presente que estos cálculos se hicieron con base en la estimación inicial $s = \$3\,000$. Si en la encuesta muestral de este año, s resulta ser más grande, la amplitud del intervalo de confianza que se obtenga será mayor que $1 000. En consecuencia, si el presupuesto lo permite, se deberá elegir una muestra de, por ejemplo, 150 para garantizar que el intervalo de confianza que se obtenga tenga una amplitud menor que $1 000.

La fórmula para determinar el tamaño de muestra necesario para estimar el total poblacional, dada una cota B del error muestral, es la siguiente.

$$n = \frac{Ns^2}{\left(\frac{B^2}{4N}\right) + s^2} \tag{22.13}$$

En el ejemplo de arriba se quería estimar el salario medio inicial con una cota del error muestral $B = 500$. Suponga que también se desea estimar el salario total de los 5 000 egresados y que la cota sea $2 millones. Con la ecuación (22.13) con $B = 2\,000\,000$ se obtendrá el tamaño de muestra necesario para obtener esa cota para el total poblacional.

*William G. Cochram, *Sampling Techniques,* 3a. ed., Wiley, 1977.

$$n = \frac{5\,000(3\,000)^2}{\dfrac{(2\,000\,000)^2}{4(5\,000)} + (3\,000)^2} = 215.31$$

Al redondear hacia arriba se ve que para obtener un intervalo de confianza de aproximadamente 95%, cuya cota sea \$2 millones, el tamaño de la muestra deberá ser de 216. Hay que hacer notar que si en esta misma encuesta se desea tener una cota de \$500 para la media poblacional y una cota de \$2 millones para el total poblacional, será necesario usar una muestra cuyo tamaño sea, por lo menos, 216. Con este tamaño de muestra se obtendrá una cota más estrecha de lo necesario para la media poblacional y la precisión mínima necesaria para el total poblacional.

El tamaño de la muestra para una estimación de la proporción poblacional, se determina con una fórmula similar a la de la media poblacional; sustituya en la ecuación (22.12) s^2 por $\bar{p}(1 - \bar{p})$, con lo que obtiene

$$n = \frac{N\bar{p}(1 - \bar{p})}{N\left(\dfrac{B^2}{4}\right) + \bar{p}(1 - \bar{p})} \qquad \textbf{(22.14)}$$

Para usar la ecuación (22.14) hay que especificar la cota B deseada y una estimación de $\bar{p}$. Si no se cuenta con una estimación de $\bar{p}$, se puede usar $\bar{p} = 0.5$; con este valor de $\bar{p}$ se garantiza que el intervalo de confianza que se obtenga tenga una cota del error muestral tan pequeña, por lo menos, como la deseada.

Ejercicios

Métodos



1. Para obtener una muestra de $n = 50$ de una población de $N = 800$, se empleó el muestreo aleatorio simple. Se halló una media muestral $\bar{x} = 215$ y una desviación estándar muestral $s = 20$.
   a. Estime la media poblacional.
   b. Estime el error estándar de la media.
   c. Obtenga un intervalo de confianza de aproximadamente 95% para la media poblacional.
2. Para obtener una muestra de $n = 80$ de una población de $N = 400$, se empleó el muestreo aleatorio simple. Se halló una media muestral $\bar{x} = 75$ y una desviación estándar muestral $s = 8$.
   a. Estime el total poblacional.
   b. Estime el error estándar del total poblacional.
   c. Obtenga un intervalo de confianza de aproximadamente 95% para el total poblacional.
3. Para obtener una muestra de $n = 100$ de una población de $N = 1\,000$, se empleó el muestreo aleatorio simple. Se halló una proporción muestral de $\bar{p} = 0.30$.
   a. Estime la proporción poblacional.
   b. Estime el error estándar de la proporción.
   c. Obtenga un intervalo de confianza de aproximadamente 95% para la proporción poblacional.
4. Se va a tomar una muestra para obtener un intervalo de confianza de aproximadamente 95% para estimar la media poblacional. La población consta de 450 elementos y en un estudio piloto se encontró $s = 70$. ¿De qué tamaño deberá ser la muestra para que la amplitud del intervalo sea 30?

Aplicaciones



5. En 1996 la Small Business Administration (SBA) concedió 771 créditos a pequeñas empresas en Carolina del Norte (*The Wall Street Journal Almanac,* 1998). Suponga que en una muestra de 50 pequeñas empresas el promedio de los créditos fue de \$149 670 y la desviación estándar de \$73 420 y que 18 de las empresas de la muestra hayan sido empresas de fabricación.
   a. Obtenga un intervalo de confianza de aproximadamente 95% para la media de los créditos.
   b. Estime un intervalo de confianza de aproximadamente 95% para el valor total de los 771 créditos en Carolina del Norte.
   c. Obtenga un intervalo de confianza de aproximadamente 95% para la proporción de créditos otorgados a empresas de fabricación.
6. En un condado de California se tienen 724 declaraciones de impuestos corporativos. El ingreso anual medio reportado es de \$161 220 con una desviación estándar de \$31 300. ¿De qué tamaño deberá ser la muestra el siguiente año para obtener un intervalo de confianza de aproximadamente 95% para el ingreso corporativo anual medio? La precisión deseada es de una amplitud de intervalo no mayor que \$5 000.



22.5 Muestreo aleatorio simple estratificado

En el **muestreo aleatorio simple estratificado**, primero se divide la población en H grupos, a los que se llama estratos. A continuación de cada estrato h se toma una muestra aleatoria simple de tamaño n_h. Los datos de las H muestras aleatorias simples se juntan para obtener una estimación del parámetro poblacional de interés como de la media, del total o de la proporción poblacionales.

Si la variabilidad dentro de cada estrato es menor que la variabilidad entre los estratos, la precisión que se obtiene con una muestra aleatoria simple estratificada puede ser muy buena (intervalos de confianza estrechos para los parámetros poblacionales). La base para la formación de los estratos depende del criterio de quien diseña la muestra. De acuerdo con la aplicación, una población puede estratificarse por departamentos, por ubicación, según la edad, el tipo de producto, el tipo de industria, la cantidad de ventas, etcétera.

Por ejemplo, suponga que el College of Business del Lakeland College desea realizar una encuesta a los egresados ese año para conocer sus salarios iniciales. En este College hay cinco áreas principales: contaduría, finanzas, sistemas de la información, marketing y administración de operaciones. De los $N = 1\,500$ estudiantes egresados ese año, $N_1 = 500$ pertenecieron a contaduría, $N_2 = 350$ a finanzas, $N_3 = 200$ a sistemas de la información, $N_4 = 300$ a marketing y $N_5 = 150$ a administración de operaciones. Los datos de análisis previos sugieren que existe mayor variabilidad entre los salarios iniciales de las distintas áreas que dentro de cada área. Por tanto, se toma una muestra aleatoria simple estratificada de $n = 180$ estudiantes; 45 de los 180 estudiantes pertenecen a contaduría ($n_1 = 45$), 40 a finanzas ($n_2 = 40$), 30 a sistemas de la información ($n_3 = 30$), 35 a marketing ($n_4 = 35$) y 30 a administración de operaciones ($n_5 = 30$).

Media poblacional

En los muestreos estratificados para obtener una estimación insesgada de la media poblacional se calcula el promedio ponderado de las medias muestrales de los estratos. Para ponderar se usa la proporción de la población que representa cada estrato. El estimador puntual que se obtiene de esta manera y que se denota $\bar{x}_{st}$, está definido como sigue.

ESTIMADOR PUNTUAL DE LA MEDIA POBLACIONAL

$$\bar{x}_{st} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \bar{x}_h \tag{22.15}$$

donde

H = número de estratos
$\bar{x}_h$ = media muestral del estrato h
N_h = número de elementos en el estrato h
N = número total de elementos en la población; $N = N_1 + N_2 + \cdots + N_H$

En el muestreo aleatorio simple estratificado, la fórmula para obtener una estimación del error estándar de la media es función de s_h, la desviación estándar muestral del estrato h.

$$s_{\bar{x}_{st}} = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h(N_h - n_h)\frac{s_h^2}{n_h}} \qquad \textbf{(22.16)}$$

Con esta fórmula el intervalo de confianza de aproximadamente 95% para estimar la media poblacional está dado por la siguiente expresión.

INTERVALO DE CONFIANZA DE APROXIMADAMENTE 95% PARA ESTIMAR LA MEDIA POBLACIONAL

$$\bar{x}_{st} \pm 2s_{\bar{x}_{st}} \qquad \textbf{(22.17)}$$

Suponga que en la encuesta realizada a los 180 egresados del College Business del Lakeland College se obtuvieron los resultados muestrales que se presentan en la tabla 22.1. Las medias muestrales en cada área o estrato son \$35 000 para contaduría, \$33 500 para finanzas, \$41 500 para sistemas de la información, \$32 000 para marketing y \$36 000 para administración de operaciones. Con estos resultados y la ecuación (22.15) se obtiene una estimación puntual para la media poblacional.

$$\bar{x}_{st} = \left(\frac{500}{1\,500}\right)(35\,000) + \left(\frac{350}{1\,500}\right)(33\,500) + \left(\frac{200}{1\,500}\right)(41\,500)$$
$$+ \left(\frac{300}{1\,500}\right)(32\,000) + \left(\frac{150}{1\,500}\right)(36\,000) = 35\,017$$

Los cálculos necesarios para estimar el error estándar se presentan en la tabla 22.2; observe que

$$\sum_{h=1}^{5} N_h(N_h - n_h)\frac{s_h^2}{n_h} = 42\,909\,037\,698$$

TABLA 22.1 ENCUESTA MUESTRAL SOBRE LOS SALARIOS INICIALES DE LOS EGRESADOS DEL LAKELAND COLLEGE

Área (h)	$\bar{x}_h$	s_h	N_h	n_h
Contaduría	\$35 000	2000	500	45
Finanzas	\$33 500	1700	350	40
Sistemas de la información	\$41 500	2300	200	30
Marketing	\$32 000	1600	300	35
Administración de operaciones	\$36 000	2250	150	30

TABLA 22.2 CÁLCULOS PARCIALES PARA LA ESTIMACIÓN DEL ERROR ESTÁNDAR DE LA MEDIA EN LA ENCUESTA MUESTRAL DEL LAKELAND COLLEGE

Área	h	$N_h(N_h - n_h)\dfrac{s_h^2}{n_h}$
Contaduría	1	$500(500 - 45)\dfrac{(2\,000)^2}{45} = 20\,222\,222\,222$
Finanzas	2	$350(350 - 40)\dfrac{(1\,700)^2}{40} = 7\,839\,125\,000$
Sistemas de la información	3	$200(200 - 30)\dfrac{(2\,300)^2}{30} = 5\,995\,333\,333$
Marketing	4	$300(300 - 35)\dfrac{(1\,600)^2}{35} = 5\,814\,857\,143$
Administración de operaciones	5	$150(150 - 30)\dfrac{(2\,250)^2}{30} = 3\,037\,500\,000$
		$42\,909\,037\,698$

$\sum_{h=1}^{5} N_h(N_h - n_h)\dfrac{s_h^2}{n_h}$ → 42 909 037 698

Por tanto,

$$s_{\bar{x}_{\text{st}}} = \sqrt{\left(\frac{1}{(1\,500)^2}\right)(42\,909\,037\,698)} = \sqrt{19\,070.68} = 138$$

De esta manera, con la ecuación (22.17), un intervalo de confianza de aproximadamente 95% para la estimación de la media poblacional es 35 017 ± 2(138) = 35.017 ± 276, es decir, \$34 741 a \$35 293.

Total poblacional

La estimación puntual del total poblacional (X) se obtiene al multiplicar N por $\bar{x}_{\text{st}}$.

ESTIMADOR PUNTUAL DEL TOTAL POBLACIONAL

$$\hat{X} = N\bar{x}_{\text{st}} \qquad \textbf{(22.18)}$$

Una estimación del error estándar de este estimador puntual es

$$s_{\hat{X}} = Ns_{\bar{x}_{\text{st}}} \qquad \textbf{(22.19)}$$

Por tanto, un intervalo de confianza de aproximadamente 95% es el dado por la expresión siguiente.

INTERVALO DE CONFIANZA DE APROXIMADAMENTE 95% PARA ESTIMAR EL TOTAL POBLACIONAL

$$N\bar{x}_{\text{st}} \pm 2s_{\hat{X}} \qquad \textbf{(22.20)}$$

Ahora suponga que el College of Business del ejemplo anterior desea estimar también el ingreso total de los 1 500 egresados con objeto de estimar su impacto en la economía. Mediante la ecuación (22.18), se obtiene una estimación insesgada del total de ingresos.

$$\hat{X} = (1\,500)35\,017 = 52\,525\,500$$

Una estimación del error estándar del total poblacional se obtiene con la ecuación (22.19).

$$s_{\hat{X}} = 1\,500(138) = 207\,000$$

En conclusión, con la ecuación (22.20), se determina que un intervalo de confianza de aproximadamente 95% para estimar los ingresos totales de los 1 500 egresados es 52 525 500 ± 2(207 000) = 52 525 500 ± 414 000, es decir \$52 111 500 a \$52 939 500.

Proporción poblacional

Una estimación insesgada de la proporción poblacional, p, cuando se emplea el muestreo aleatorio simple estratificado, es un promedio ponderado de las proporciones de cada estrato. Para ponderar se usan las fracciones de la población que corresponden a cada estrato. El estimador puntual que se obtiene, denotado $\bar{p}_{st}$, se define como sigue.

ESTIMADOR PUNTUAL DE LA PROPORCIÓN POBLACIONAL

$$\bar{p}_{st} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \bar{p}_h \qquad \textbf{(22.21)}$$

donde

H = número de estratos
$\bar{p}_h$ = proporción muestral del estrato h
N_h = número de elementos de la población que pertenecen al estrato h
N = número total de elementos en la población: $N = N_1 + N_2 + \cdots + N_H$

Una estimación del error estándar de $\bar{p}_{st}$ es la dada por

$$s_{\bar{p}_{st}} = \sqrt{\frac{1}{N^2} \sum_{h=1}^{H} N_h(N_h - n_h)\left[\frac{\bar{p}_h(1 - \bar{p}_h)}{n_h - 1}\right]} \qquad \textbf{(22.22)}$$

Por tanto, un intervalo de confianza de aproximadamente 95% para estimar la proporción poblacional es el dado por la expresión siguiente.

INTERVALO DE CONFIANZA DE APROXIMADAMENTE 95% PARA ESTIMAR LA PROPORCIÓN POBLACIONAL

$$\bar{p}_{st} \pm 2s_{\bar{p}_{st}} \qquad \textbf{(22.23)}$$

Suponga que en este ejemplo se desea conocer la proporción de recién egresados que obtuvo un salario inicial de \$36 000 o más. En los resultados de la encuesta muestral de los 180 recién egresados se observa que 63 de ellos obtuvieron un salario inicial de \$36 000 o más y que 16 de los

63 corresponden a contaduría, 3 a finanzas, 29 a sistemas de la información, 0 a marketing y 15 a administración de operaciones.

Con la ecuación (22.21) se calcula la proporción de egresados que obtuvo un salario inicial de $36 000 o más.

$$\bar{p}_{\text{st}} = \left(\frac{500}{1\,500}\right)\left(\frac{16}{45}\right) + \left(\frac{350}{1\,500}\right)\left(\frac{3}{40}\right) + \left(\frac{200}{1\,500}\right)\left(\frac{29}{30}\right) + \left(\frac{300}{1\,500}\right)\left(\frac{0}{35}\right) + \left(\frac{150}{1\,500}\right)\left(\frac{15}{30}\right)$$
$$= 0.3149$$

Los cálculos necesarios para estimar el error estándar se muestran en la tabla 22.3; observe que

$$\sum_{h=1}^{5} N_h(N_h - n_h)\left[\frac{\bar{p}_h(1-\bar{p}_h)}{n_h - 1}\right] = 1570.6913$$

Por tanto,

$$s_{\bar{p}_{\text{st}}} = \sqrt{\frac{1}{(1\,500)^2}(1570.6913)}$$
$$= 0.0264$$

Con la expresión (22.23), se encuentra que un intervalo de confianza de aproximadamente 95% para la proporción de egresados que tienen un salario inicial de $36 000 o más es 0.3149 ± 2(0.0264) = 0.3149 ± 0.0528, es decir 0.2621 a 0.3677.

Determinación del tamaño de la muestra

Cuando se emplea un muestreo aleatorio simple estratificado, la elección del tamaño de la muestra se entiende como un proceso de dos pasos. Primer paso, se elige un tamaño total n para la muestra. Segundo paso, se decide cuántas unidades muestrales tomar de cada estrato. Otra alternativa es decidir primero de qué tamaño se tomará la muestra de cada estrato y después sumar

TABLA 22.3 CÁLCULOS PARCIALES PARA LA ESTIMACIÓN DEL ERROR ESTÁNDAR DE $\bar{p}_{\text{st}}$ EN LA ENCUESTA MUESTRAL DE LOS ESTUDIANTES DEL LAKELAND

Área	h	$N_h(N_h - n_h)\left[\frac{\bar{p}_h(1-\bar{p}_h)}{n_h - 1}\right]$
Contaduría	1	$500(500 - 45)\left[\frac{(16/45)(29/45)}{45 - 1}\right] = 1\,184.7363$
Finanzas	2	$350(350 - 40)\left[\frac{(3/40)(37/40)}{40 - 1}\right] = 193.0048$
Sistemas de la información	3	$200(200 - 30)\left[\frac{(29/30)(1/30)}{30 - 1}\right] = 37.7778$
Marketing	4	$300(300 - 35)\left[\frac{(0/35)(35/35)}{35 - 1}\right] = 0.0000$
Administración de operaciones	5	$150(150 - 30)\left[\frac{(15/30)(15/30)}{30 - 1}\right] = 155.1724$
		1 570.6913
	$\sum_{h=1}^{5} N_h(N_h - n_h)\left[\frac{\bar{p}_h(1-\bar{p}_h)}{n_h - 1}\right]$	

los tamaños muestrales de los estratos para obtener el tamaño total de la muestra. Con frecuencia interesa obtener estimaciones de la media, el total y la proporción de cada estrato; por tanto se suele emplear una combinación de estos dos métodos. Se determina un tamaño general n de la muestra y una asignación con la que se obtenga la precisión necesaria para los parámetros poblacionales generales de interés. Después, si los tamaños de las muestras de algunos de los estratos no son lo suficientemente grandes para obtener la precisión necesaria en las estimaciones para el estrato, se aumentan los tamaños de las muestras de esos estratos. En esta subsección se verán algunos de los aspectos relacionados con la asignación de toda la muestra a los diferentes estratos y se presentará un método para elegir el tamaño total de la muestra y hacer la asignación.

La asignación consiste en decidir qué fracción del total de la muestra le será asignada a cada estrato. Esta fracción determina cuán grande será la muestra aleatoria simple de cada estrato. Los factores más importantes para hacer la asignación son los siguientes:

1. El número de elementos en cada estrato.
2. La varianza de los elementos dentro de cada estrato.
3. El costo de la selección de los elementos de cada estrato.

En general, se asignan muestras más grandes a los estratos más grandes y a los estratos que tienen mayor varianza. En sentido inverso, para obtener la mayor información por un costo dado, a los estratos que tienen un costo por unidad muestreada mayor se les asignan muestras más pequeñas.

Las varianzas de cada uno de los estratos suelen ser muy diferentes. Por ejemplo, suponga que en un determinado estudio se desee determinar la cantidad media de empleados por edificio; como la variabilidad será mayor en un estrato que tenga edificios grandes que en un estrato que tenga edificios pequeños, en tales estratos se tomará una muestra proporcionalmente mayor. El costo de la selección puede ser una consideración importante cuando el entrevistador tiene que hacer recorridos significativos entre las unidades muestrales de unos estratos, pero no entre las de otros; esta situación suele surgir cuando algunos de los estratos comprenden zonas rurales y otras ciudades.

En muchas encuestas el costo por unidad de muestreo es aproximadamente el mismo en todos los estratos (por ejemplo, en las encuestas por correo o por teléfono); en tales casos el costo del muestreo puede ignorarse al hacer la asignación. Aquí se presentan las fórmulas apropiadas para elegir el tamaño de la muestra y para hacer la asignación en tales casos. En los libros más avanzados sobre muestreo se proporcionan las fórmulas para el caso en el que los costos de muestreo varían significativamente entre los estratos. Las fórmulas que se presentan en esta sección minimizan el costo total de muestreo dado un nivel de precisión. Este método conocido como *asignación de Neyman* asigna el total de la muestra n a los diversos estratos como sigue.

$$n_h = n\left(\frac{N_h s_h}{\sum_{h=1}^{H} N_h s_h}\right) \tag{22.24}$$

La ecuación (22.24) indica que el número de unidades asignadas a un estrato aumenta con el tamaño del estrato y la desviación estándar. Observe que para hacer esta asignación primero se necesita determinar el tamaño total n de la muestra. Dada una determinada precisión B, para elegir el tamaño de la muestra para la determinación de la media poblacional y del total poblacional se emplean las fórmulas siguientes.

TAMAÑO DE LA MUESTRA PARA ESTIMAR LA MEDIA POBLACIONAL

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{N^2\left(\frac{B^2}{4}\right) + \sum_{h=1}^{H} N_h s_h^2} \tag{22.25}$$

TAMAÑO DE LA MUESTRA PARA ESTIMAR EL TOTAL POBLACIONAL

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{\frac{B^2}{4} + \sum_{h=1}^{H} N_h s_h^2} \quad \textbf{(22.26)}$$

Por ejemplo, suponga que un vendedor de Chevrolet desea hacer una encuesta a los clientes que compran un Corvette o un Cavalier o un Geo Prizm para obtener información que él considera puede ser útil para la publicidad futura. Suponga que este vendedor desea estimar el ingreso medio mensual de estos clientes y que la cota del error muestral sea $100. Los 600 clientes de este vendedor se dividen en tres estratos: 100 que poseen un Corvette, 200 que poseen un Geo Prizm y 300 que poseen un Cavalier. Para estimar la desviación estándar de cada estrato se empleó una encuesta piloto; los resultados son $s_1 = \$1\,300$, $s_2 = \$900$ y $s_3 = \$500$ para los poseedores de un Corvette, un Geo Prizm y un Cavalier, respectivamente.

El primer paso para elegir el tamaño de la muestra es usar la ecuación (22.25) para determinar el tamaño total de la muestra que se necesita para tener una cota $B = \$100$ en la estimación de la media poblacional. Primero, se calcula

$$\sum_{h=1}^{3} N_h s_h = 100(1\,300) + 200(900) + 300(500) = 460\,000$$

Después se calcula

$$\sum_{h=1}^{3} N_h s_h^2 = 100(1\,300)^2 + 200(900)^2 + 300(500)^2 = 406\,000\,000$$

Al sustituir estos valores en la ecuación (22.25), se obtiene el tamaño total de la muestra que se requiere para que la cota del error muestral sea $B = \$100$.

$$n = \frac{(460\,000)^2}{\frac{(600)^2(100)^2}{4} + 406\,000\,000} = 162$$

Por tanto, el tamaño total de la muestra para obtener la precisión deseada es 162. Para la asignación del total de la muestra a los tres estratos se emplea la ecuación (22.24).

$$n_1 = 162\left(\frac{100(1\,300)}{460\,000}\right) = 46$$

$$n_2 = 162\left(\frac{200(900)}{460\,000}\right) = 63$$

$$n_3 = 162\left(\frac{300(500)}{460\,000}\right) = 53$$

De manera que la recomendación será: 46 propietarios de Corvette, 62 de Geo Prizm y 53 de Cavalier que hacen un tamaño total de la muestra de 162 clientes.

Para determinar el tamaño de la muestra necesario para la estimación de la proporción poblacional, simplemente se sustituye en la ecuación (22.25) s_h por $\sqrt{\bar{p}_h(1-\bar{p}_h)}$; el resultado es

$$n = \frac{\left(\sum_{h=1}^{H} N_h \sqrt{\bar{p}_h(1-\bar{p}_h)}\right)^2}{N^2\left(\frac{B^2}{4}\right) + \sum_{h=1}^{H} N_h \bar{p}_h(1-\bar{p}_h)} \quad \textbf{(22.27)}$$

Una vez que se ha determinado el tamaño total de la muestra necesario para la estimación de la proporción poblacional, la asignación a los varios estratos se hace otra vez con la ecuación (22.25) al sustituir s_h por $\sqrt{\bar{p}_h(1-\bar{p}_h)}$.

NOTAS Y COMENTARIOS

1. Una ventaja del muestreo aleatorio simple estratificado es que automáticamente, como consecuencia del procedimiento de muestreo, se obtienen estimaciones de los parámetros poblacionales de cada estrato. Por ejemplo, además de obtener una estimación del salario inicial promedio en el problema de los egresados de administración, se obtuvo también una estimación del salario inicial promedio de los egresados de cada área. Como cada una de las estimaciones de los salarios iniciales se hizo con base en una muestra aleatoria simple de cada estrato, se puede emplear el procedimiento para obtener un intervalo de confianza cuando se toma una muestra aleatoria simple (véase la ecuación (22.4)) para calcular una estimación de la media de cada estrato mediante un intervalo de confianza de aproximadamente 95% para estimar la media de cada estrato. De manera similar, se obtienen intervalos de confianza para estimar el total poblacional y la proporción poblacional de cada estrato con las ecuaciones (22.8) y (22.10), respectivamente.
2. Otro tipo de asignación que se usa cuando se hace un muestreo aleatorio simple estratificado es la *asignación proporcional.* Con este método el tamaño de la muestra asignada a cada estrato está dado por la fórmula siguiente.

$$n_h = n\left(\frac{N_h}{N}\right) \quad \textbf{(22.28)}$$

La asignación proporcional debe usarse cuando las varianzas de los estratos son todas aproximadamente iguales y el costo por unidad muestreada es casi el mismo en todos los estratos. En los casos en que las varianzas de los estratos son iguales, la asignación proporcional y el procedimiento de Neyman dan las mismas asignaciones.

Ejercicios

Métodos



7. Los resultados obtenidos de una muestra aleatoria simple estratificada fueron los siguientes.

Estrato (h)	$\bar{x}_h$	s_h	$\bar{p}_h$	N_h	n_h
1	138	30	0.50	200	20
2	103	25	0.78	250	30
3	210	50	0.21	100	25

a. Proporcione una estimación de la media poblacional de cada estrato.
b. Dé un intervalo de confianza de aproximadamente 95% para la media poblacional de cada estrato.
c. Encuentre un intervalo de confianza de aproximadamente 95% para la media poblacional de toda la población.

8. Reconsidere los resultados muestrales del ejercicio 7.
   a. Encuentre una estimación del total poblacional de cada estrato.
   b. Dé una estimación puntual del total de los 550 elementos de la población.
   c. Proporcione un intervalo de confianza de aproximadamente 95% para el total poblacional.
9. Regrese a los resultados muestrales del ejercicio 7.
   a. Proporcione un intervalo de confianza de aproximadamente 95% para la proporción de cada estrato.
   b. Dé una estimación puntual de la proporción poblacional para los 550 elementos de la población.
   c. Estime el error estándar de la proporción poblacional.
   d. Encuentre un intervalo de aproximadamente 95% para la proporción poblacional.
10. Una población fue dividida en tres estratos $N_1 = 300$, $N_2 = 600$ y $N_3 = 500$. De una encuesta anterior se tienen las estimaciones siguientes de la desviación estándar en cada uno de los estratos: $s_1 = 150$, $s_2 = 75$ y $s_3 = 100$.
    a. Suponga que se necesita una estimación de la media poblacional con una cota del error de estimación de $B = 20$. ¿De qué tamaño deberá ser la muestra? ¿Cuántos elementos deberán ser tomados de cada estrato?
    b. Admita que se requiere que la cota sea $B = 10$ ¿De qué tamaño deberá ser la muestra? ¿Cuántos elementos deberán ser tomados de cada estrato?
    c. Suponga que se quiere tener una estimación del total poblacional y que la cota sea $B = 15\,000$. ¿De qué tamaño deberá ser la muestra? ¿Cuántos elementos deberán ser tomados de cada estrato?

Aplicaciones

11. Una cadena de farmacias tiene tiendas en cuatro ciudades: 38 tiendas en Indianápolis, 45 en Louisville, 80 en St. Louis y 70 en Memphis. Las ventas en las cuatro ciudades varían considerablemente debido a la competencia. De una encuesta muestral se tienen los datos siguientes (en miles de dólares). Cada una de las ciudades se consideró como un estrato y se tomó una muestra aleatoria simple estratificada.

Indianápolis	Louisville	St. Louis	Memphis
50.3	48.7	16.7	14.7
41.2	59.8	38.4	88.3
15.7	28.9	51.6	94.2
22.5	36.5	42.7	76.8
26.7	89.8	45.0	35.1
20.8	96.0	59.7	48.2
	77.2	80.0	57.9
	81.3	27.6	18.8
			22.0
			74.3

   a. Estime la media de las ventas en cada ciudad (estrato).
   b. Obtenga un intervalo de confianza de aproximadamente 95% para la media de las ventas en cada ciudad.
   c. Estime la proporción de farmacias cuyas ventas son de $50 000 o más.
   d. Obtenga un intervalo de confianza de aproximadamente 95% para la proporción de farmacias cuyas ventas son de $50 000 o más.
12. Reconsidere los resultados de la encuesta muestral del ejercicio 11.
    a. Estime el total poblacional de las ventas en St. Louis.
    b. Valore el total poblacional de las ventas en Indianápolis.
    c. Obtenga un intervalo de confianza de aproximadamente 95% para la media de las ventas de la cadena de farmacias.

d. Encuentre un intervalo de confianza de aproximadamente 95% para el total de las ventas de la cadena de farmacias.

13. Una empresa de contadores tiene clientes en la industria bancaria, en la de seguros y en la de corretaje: $N_1 = 50$ bancos, $N_2 = 38$ empresas de seguros y $N_2 = 35$ empresas de corretaje. Esta empresa ha contratado a una empresa que se dedica al marketing para que realice una encuesta entre sus clientes en estas tres industrias. En la encuesta se harán diversas preguntas relacionadas, tanto con el negocio de los clientes como con su satisfacción con el servicio que reciben de la empresa de contadores. Suponga que se desea un intervalo de confianza de aproximadamente 95% para estimar el número promedio de empleados en los 123 clientes con una cota en el error de estimación $B = 30$.
   a. Suponga que en un estudio piloto se encuentra $s_1 = 80$, $s_2 = 150$ y $s_3 = 45$. Determine el tamaño total de la muestra y explique los tamaños de las muestras asignados a los tres estratos.
   b. Suponga que se duda del estudio piloto y que para determinar el tamaño de la muestra se decide suponer que todos los estratos tienen la misma desviación estándar de 100. Determine el tamaño de la muestra y diga cuántos elementos deben tomarse de cada estrato.



Muestreo por conglomerados

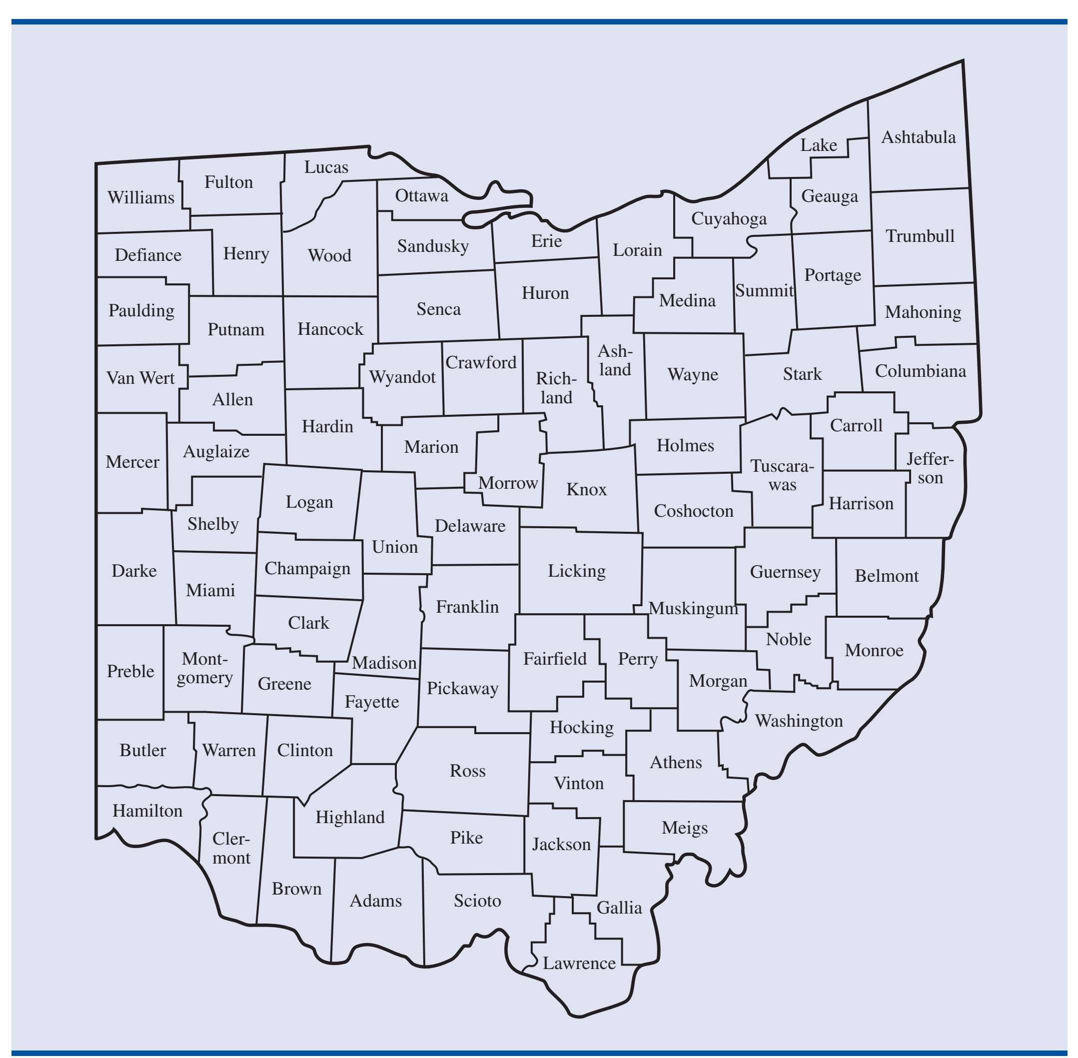
En el **muestreo por conglomerados** se requiere que la población se divida en N grupos de elementos llamados conglomerados, de manera que cada elemento de la población pertenezca a uno y sólo un conglomerado. Por ejemplo, suponga que se quiere investigar a los votantes registrados del estado de Ohio. Un método de muestreo puede ser elaborar un marco con todos los votantes registrados del estado de Ohio y, de este marco, tomar una muestra aleatoria simple. De manera alternativa, para el muestreo por aglomerados se define el marco como la lista de los $N = 88$ condados del estado (véase figura 22.1). Con este método, cada condado o conglomerado consiste en un grupo de votantes registrados y cada votante registrado del estado pertenece a uno y sólo un conglomerado.

Suponga que de los 88 condados se toma una muestra aleatoria simple de $n = 12$ condados. Ahora se pueden recolectar los datos de *todos* los votantes registrados de cada uno de los 12 conglomerados, método que se conoce como *muestreo por conglomerados en una sola etapa* o se puede tomar una muestra aleatoria simple de los votantes registrados en cada uno de los 12 conglomerados muestreados, método que se conoce como *muestreo por conglomerados en dos etapas*. En cualquier caso existen fórmulas para usar los resultados muestrales en la obtención de estimaciones puntuales o por intervalo de estimación de parámetros poblacionales como la media, el total o la proporción poblacionales. En este capítulo se considerará únicamente el muestreo por conglomerados en una sola etapa; en libros más avanzados sobre muestreo se encuentra la información sobre el muestreo por conglomerados en dos etapas.

El muestreo estratificado y el muestreo por conglomerados son similares, ya que en los dos se divide a la población en grupos de elementos. Sin embargo, las razones para elegir el muestreo por conglomerados son diferentes de las razones para elegir el muestreo estratificado. Con el muestreo por conglomerados se obtienen mejores resultados cuando los elementos dentro de los conglomerados son heterogéneos (no son parecidos). Lo ideal es que los conglomerados sean una versión a pequeña escala de la población. Cuando es así, al muestrear una pequeña cantidad de conglomerados se obtiene una buena información de las características de toda la población.

Una de las principales aplicaciones del muestreo por conglomerados es el muestreo de áreas, en donde los conglomerados son condados, poblaciones, manzanas u otras secciones geográficamente bien definidas de la población. Como sólo se recolectan datos de una muestra o conglomerado de toda un área geográfica y como los elementos dentro de un conglomerado se suelen encontrar cerca, uno de otros, cuando un recolector de datos o entrevistador se envía a una unidad muestreada se obtiene un considerable ahorro de tiempo y costos. Por tanto, aun cuando se requiera que el tamaño total de la muestra sea grande, el muestreo por conglomerados puede resultar menos costoso que el muestreo aleatorio simple o el muestreo aleatorio simple estratificado. Además,

FIGURA 22.1 CONDADOS DEL ESTADO DE OHIO USADOS COMO CONGLOMERADOS DE VOTANTES REGISTRADOS



No es necesario hacer una lista de todos los elementos de la población. En el muestreo por conglomerados sólo se necesita una lista de los elementos en los conglomerados muestreados.

el muestreo por conglomerados minimiza el tiempo y el costo de la elaboración de un marco o lista de los elementos a ser muestreados, dado que en el muestreo por conglomerados no se necesita una lista de todos los elementos de la población. Sólo se necesita una lista de los elementos de los conglomerados muestreados.

Con el fin de ilustrar el muestreo por conglomerados se verá una encuesta realizada por la CPA Society (Certified Public Account Society) de los 12 000 integrantes en servicio en un determinado estado. Como parte de la encuesta, la CPA Society recolectó información sobre el ingreso, el género y sobre factores relacionados con el estilo de vida de los contadores. Dado que para obtener toda la información deseada era necesario realizar entrevistas, para minimizar los gastos de desplazamiento y de las entrevistas, la CPA Society realizó un muestreo por conglomerados. El marco consistió en todas las empresas con contadores registradas en el estado. Suponga que el número de empresas registradas en el estado haya sido $N = 1\,000$ y que se haya tomado una muestra aleatoria simple de $n = 10$ empresas.

Para las fórmulas que se requieren en el muestreo por conglomerados para la obtención de intervalos de confianza de aproximadamente 95% para estimar la media, el total o la proporción poblacionales, se usará la notación siguiente.

N = número de conglomerados en la población
n = número de conglomerados tomados para la muestra
M_i = número de elementos en el conglomerado i
M = número de elementos en la población; $M = M_1 + M_2 + \cdots + M_N$
$\bar{M} = M/N$ = número promedio de elementos en un conglomerado
x_i = número total de observaciones en el conglomerado i
a_i = número de observaciones con una determinada característica en el conglomerado i

En el caso de la encuesta muestral de la CPA Society, la información que se tiene es la siguiente.

$$N = 1\,000$$
$$n = 10$$
$$M = 12\,000$$
$$\bar{M} = 12\,000/1\,000 = 12$$

En la tabla 22.4 se dan los valores de M_i y x_i correspondientes a cada uno de los conglomerados muestreados, así como el número de mujeres de la CPA Society en las empresas muestreadas.

Media poblacional

El estimador puntual de la media poblacional en un muestreo por conglomerados está dado por la fórmula siguiente.

ESTIMADOR PUNTUAL DE LA MEDIA POBLACIONAL

$$\bar{x}_c = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i} \qquad \textbf{(22.29)}$$

TABLA 22.4 RESULTADOS DE LA ENCUESTA MUESTRAL SOBRE LOS CPA

Empresa (i)	CPA (M_i)	Salario total (miles de $) en la empresa i (x_i)	Mujeres CPA (a_i)
1	8	384	2
2	25	1350	8
3	4	148	0
4	17	857	6
5	7	296	1
6	3	131	2
7	15	761	2
8	4	176	0
9	12	577	5
10	33	1880	9
Totales	128	6560	35

Una estimación del error estándar de este estimador puntual es

$$s_{\bar{x}_c} = \sqrt{\left(\frac{N - n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(x_i - \bar{x}_c M_i)^2}{n - 1}} \quad \textbf{(22.30)}$$

Por tanto, con la expresión siguiente se obtiene un intervalo de aproximadamente 95% como estimación de la media poblacional.

INTERVALO DE APROXIMADAMENTE 95% COMO ESTIMACIÓN DE LA MEDIA POBLACIONAL

$$\bar{x}_c \pm 2s_{\bar{x}_c} \quad \textbf{(22.31)}$$

Al emplear los datos de la tabla 22.4 se obtiene una estimación del salario medio de los contadores públicos certificados.

$$\bar{x}_c = \frac{6\,560}{128} = 51.250$$

Los datos de la tabla 22.4, dados en miles de dólares, indican que una estimación del salario medio de los contadores públicos certificados del estado es $51 250.

En la tabla 22.5 se presenta parte de los cálculos que se necesitan para estimar el error estándar, observe que

$$\sum_{i=1}^{n}(x_i - \bar{x}_c M_i)^2 = 51{,}281.378$$

Por tanto,

$$s_{\bar{x}_c} = \sqrt{\left[\frac{1\,000 - 10}{(1\,000)(10)(12)^2}\right]\frac{51\,281.378}{10 - 1}} = 1.979$$

Por tanto, el error estándar es $1 979. Con la expresión (22.31) se encuentra que un intervalo de confianza de aproximadamente 95% para la estimación del salario anual medio es 51 250 ± 2(1 979) = 51 250 ± 3 958, es decir, $47 292 a $55 208.

Total poblacional

El estimador puntual para el total poblacional se obtiene multiplicando M por $\bar{x}_c$.

ESTIMADOR PUNTUAL PARA EL TOTAL POBLACIONAL

$$\hat{X} = M\bar{x}_c \quad \textbf{(22.32)}$$

Una estimación del error estándar de este estimador es

$$s_{\hat{X}} = Ms_{\bar{x}_c} \quad \textbf{(22.33)}$$

TABLA 22.5 CÁLCULOS PARCIALES PARA LA ESTIMACIÓN DEL ERROR ESTÁNDAR DE LA MEDIA EN LA ENCUESTA MUESTRAL DE CPA

Empresa (i)	M_i	x_i	$(x_i - 51.250M_i)^2$
1	8	384	$[384 - 51.250(8)]^2 =$ 676.000
2	25	1350	$[1350 - 51.250(25)]^2 =$ 4 726.563
3	4	148	$[148 - 51.250(4)]^2 =$ 3 249.000
4	17	857	$[857 - 51.250(17)]^2 =$ 203.063
5	7	296	$[296 - 51.250(7)]^2 =$ 3 937.563
6	3	131	$[131 - 51.250(3)]^2 =$ 517.563
7	15	761	$[761 - 51.250(15)]^2 =$ 60.063
8	4	176	$[176 - 51.250(4)]^2 =$ 841.000
9	12	577	$[577 - 51.250(12)]^2 =$ 1 444.000
10	33	1880	$[1880 - 51.250(33)]^2 =$ 35 626.563
Totales	128	6560	51 281.378

$\sum_{i=1}^{n}(x_i - \bar{x}_c M_i)^2$ → 51 281.378

Por tanto, un intervalo de confianza de aproximadamente 95% para estimar el total poblacional está dado por la expresión siguiente.

INTERVALO DE CONFIANZA DE APROXIMADAMENTE 95% PARA EL TOTAL POBLACIONAL

$$M\bar{x}_c \pm 2s_{\hat{X}} \tag{22.34}$$

En la encuesta muestral de la CPA.

$$\hat{X} = M\bar{x}_c = 12\,000(51\,250) = \$615\,000\,000$$
$$s_{\hat{X}} = Ms_{\bar{x}_c} = 12\,000(1\,979) = \$23\,748\,000$$

En consecuencia, con la expresión (22.34), se encuentra que un intervalo de confianza de aproximadamente 95% es \$615 000 000 ± 2(\$23 784 000) = \$615 000 000 ± \$47 496 000, es decir, \$567 504 000 a \$662 496 000.

Proporción poblacional

A continuación se da el estimador puntual para la proporción poblacional cuando se hace un muestreo por conglomerados.

ESTIMADOR PUNTUAL PARA LA PROPORCIÓN POBLACIONAL

$$\bar{p}_c = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i} \tag{22.35}$$

donde

a_i = número de elementos, con la característica de interés, en el conglomerado i.

Una estimación de la desviación estándar de este estimador puntual es

$$s_{\bar{p}_c} = \sqrt{\left(\frac{N - n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(a_i - \bar{p}_c M_i)^2}{n - 1}} \quad \textbf{(22.36)}$$

Entonces un intervalo de confianza de aproximadamente 95% para estimar la proporción poblacional es el dado por la expresión siguiente.

INTERVALO DE CONFIANZA DE APROXIMADAMENTE 95% PARA ESTIMAR LA PROPORCIÓN POBLACIONAL

$$\bar{p}_c \pm 2 s_{\bar{p}_c} \quad \textbf{(22.37)}$$

En el caso de la encuesta muestral de los contadores certificados con la ecuación (22.35) y los datos de la tabla 22.4 se obtiene una estimación de la proporción de contadores certificados que son mujeres.

$$\bar{p}_c = \frac{2 + 8 + \cdots + 9}{8 + 25 + \cdots + 33} = \frac{35}{128} = 0.2734$$

En la tabla 22.6 se presenta parte de los cálculos necesarios para estimar el error estándar; observe que

$$\sum_{i=1}^{n}(a_i - \bar{p}_c M_i)^2 = 15.2098$$

Por tanto,

$$s_{\bar{p}_c} = \sqrt{\left[\frac{1\,000 - 10}{(1\,000)(10)(12)^2}\right]\frac{15.2098}{10 - 1}} = 0.0341$$

De esta manera, con la expresión (22.37) se halla que un intervalo de confianza de aproximadamente 95% para la proporción de mujeres que son contadoras certificadas es 0.2734 ± 2(0.0341) = 0.2734 ± 0.0682, es decir, 0.2052 a 0.3416.

Determinación del tamaño de la muestra

Una vez formados los conglomerados, lo primero para determinar el tamaño de la muestra es elegir el número n de conglomerados. Este procedimiento en el caso del muestreo por conglomerados es parecido al procedimiento empleado con los otros métodos de muestreo. El nivel de precisión se especifica al elegir un valor para B, la cota del error muestral. Después se elabora la fórmula para obtener el valor de n que permitirá lograr la precisión deseada.

Para decidir cuántos conglomerados incluir en la muestra, los factores decisivos son el tamaño promedio de los conglomerados y la varianza entre los conglomerados. Si los conglomerados son parecidos, la varianza entre ellos será pequeña y el número de conglomerados que se muestree puede ser pequeña. Las fórmulas para determinar exactamente el tamaño de la muestra se encuentran en libros más avanzados sobre muestreo.

TABLA 22.6 PARTE DE LOS CÁLCULOS PARA LA ESTIMACIÓN DEL ERROR ESTÁNDAR DE $\bar{p}_c$ EN EL ESTUDIO MUESTRAL DE LA CPA, DONDE $\bar{p}_c = 0.2734$

Empresa (i)	M_i	a_i	$(a_i - 0.2734M_i)^2$
1	8	2	$[2 - 0.2734(8)]^2 =$ 0.0350
2	25	8	$[8 - 0.2734(25)]^2 =$ 1.3572
3	4	0	$[0 - 0.2734(4)]^2 =$ 1.1960
4	17	6	$[6 - 0.2734(17)]^2 =$ 1.8284
5	7	1	$[1 - 0.2734(7)]^2 =$ 0.8350
6	3	2	$[2 - 0.2734(3)]^2 =$ 1.3919
7	15	2	$[2 - 0.2734(15)]^2 =$ 4.4142
8	4	0	$[0 - 0.2734(4)]^2 =$ 1.1960
9	12	5	$[5 - 0.2734(12)]^2 =$ 2.9556
10	33	9	$[9 - 0.2734(33)]^2 =$ 0.0005
Totales	128	35	15.2098

$$\sum_{i=1}^{n}(a_i - \bar{p}_c M_i)^2$$

Ejercicios

Métodos



14. De una población que tiene $N = 25$ conglomerados y $M = 300$ elementos se va a tomar una muestra de cuatro conglomerados. En la tabla siguiente se presentan los valores de M_i, x_i y a_i en cada conglomerado.

Conglomerado (i)	M_i	x_i	a_i
1	7	95	1
2	18	325	6
3	15	190	6
4	10	140	2
Totales	50	750	15

a. Dé estimaciones puntuales de la media, del total y de la proporción poblacionales.
b. Estime el error estándar de las estimaciones del inciso a.
c. Obtenga un intervalo de confianza de aproximadamente 95% para la media poblacional.
d. Proporcione un intervalo de confianza de aproximadamente 95% para el total poblacional.
e. Obtenga un intervalo de confianza de aproximadamente 95% para la proporción poblacional.

15. De una población que tiene $N = 30$ conglomerados y $M = 600$ elementos se va a tomar una muestra de cuatro conglomerados. En la tabla siguiente se presentan los valores de M_i, x_i y a_i en cada conglomerado

Conglomerado (i)	M_i	x_i	a_i
1	35	3 500	3
2	15	965	0
3	12	960	1
4	23	2 070	4
5	20	1 100	3
6	25	1 805	2
Totales	130	10 400	13

a. Dé estimaciones puntuales de la media, del total y de la proporción poblacionales.
b. Obtenga un intervalo de confianza de aproximadamente 95% para la media poblacional.
c. Proporcione un intervalo de confianza de aproximadamente 95% para el total poblacional.
d. Obtenga un intervalo de confianza de aproximadamente 95% para la proporción poblacional.

Aplicaciones

16. Una empresa de servicio público realiza una encuesta a los ingenieros mecánicos para conocer los factores que influyen en la elección del equipo de calefacción, ventilación y aire acondicionado (HVAC, por sus siglas en inglés) en los nuevos edificios comerciales. En el área de acción de esta empresa de servicio público hay 120 empresas relacionadas con el diseño de sistemas HVAC. La idea es hacer un muestreo por conglomerados en el que cada empresa represente un conglomerado. Todos los ingenieros mecánicos de cada empresa que se tome en la muestra serán entrevistados. Se cree que en las 120 empresas están empleados aproximadamente 500 ingenieros mecánicos. Se toma una muestra de 10 empresas. Entre otras cosas se anota la edad de cada entrevistado y si estudió en la universidad local.

Conglomerado (i)	M_i	Total de las edades de los entrevistados	Número que estudió en la universidad local
1	12	520	8
2	1	33	0
3	2	70	1
4	1	29	1
5	6	270	3
6	3	129	2
7	2	102	0
8	1	48	1
9	9	337	7
10	13	462	12
Totales	50	2000	35

a. Estime la edad promedio de los ingenieros mecánicos que trabajan en esta actividad.
b. Halle la proporción de ingenieros mecánicos en esta actividad que estudió en la universidad local.
c. Dé un intervalo de confianza de aproximadamente 95% para la edad promedio de los ingenieros mecánicos que trabajan en esta actividad.
d. Proporcione un intervalo de confianza de aproximadamente 95% para la proporción de ingenieros mecánicos en el área de acción de la empresa pública que estudiaron en la universidad local.

17. Una empresa inmobiliaria nacional acaba de adquirir una empresa más pequeña que cuenta con 150 oficinas y 6 000 agentes en Los Ángeles y en otros sitios del sur de California. La empresa nacional realizó una encuesta muestral para conocer las actitudes y otras características de sus nuevos empleados. En una muestra de ocho oficinas, todos los empleados llenaron los cuestionarios. A continuación se presentan los resultados de esta encuesta en las ocho oficinas.

Oficina	Agentes	Edad promedio	Título universitario	Agentes del sexo masculino
1	17	37	3	4
2	35	32	14	12
3	26	36	8	7
4	66	30	38	28
5	43	41	18	12
6	12	52	2	6
7	48	35	20	17
8	57	44	25	26

a. Estime la edad promedio de los agentes.
b. Estime la proporción de agentes que tienen un título universitario y la proporción de agentes del sexo masculino.
c. Dé un intervalo de confianza de aproximadamente 95% para la edad promedio de los agentes.
d. Dé un intervalo de confianza de aproximadamente 95% para la proporción de agentes que tiene un título universitario.
e. Dé un intervalo de confianza de aproximadamente 95% para la proporción de agentes del sexo masculino.



Muestreo sistemático

El **muestreo sistemático** suele emplearse como alternativa al muestreo aleatorio simple. En algunas situaciones, en especial cuando las poblaciones son grandes se lleva mucho tiempo tomar una muestra aleatoria simple en la que primero hay que hallar un número aleatorio y después contar o buscar en el marco el elemento correspondiente. En tales casos el muestreo sistemático es una alternativa al muestreo aleatorio simple. Por ejemplo, si se quiere una muestra de tamaño 50 de una población que contiene 5 000 elementos, se toma en la muestra un elemento cada 5 000/50 = 100 elementos de la población. En este caso un muestreo sistemático consistirá en elegir en forma aleatoria uno de los primeros 100 elementos del marco. Los demás elementos para la muestra se encuentran contando a partir del primer elemento tomado en la muestra hasta encontrar el centésimo elemento que sigue en el marco. En efecto, la muestra de 50 elementos se obtiene avanzando sistemáticamente a través de la población y tomando el centésimo elemento después del último elemento tomado en la muestra. De esta manera suele ser mas fácil tomar la muestra de 50 elementos de lo que sería si se usara un muestreo aleatorio simple. El primer elemento se elige aleatoriamente, lo que permite suponer que una muestra sistemática tiene las propiedades de una muestra aleatoria simple. Esta suposición suele ser correcta cuando el marco es un ordenamiento aleatorio de los elementos de la población.

Resumen

Se le presentó una breve introducción al campo de muestreo de encuestas. El propósito del muestreo de encuestas es recolectar datos con el objeto de hacer estimaciones de parámetros poblacionales como la media, el total o la proporción poblacional. El muestreo de encuestas es comparable con la realización de experimentos para generar datos. Cuando se usa el muestreo de encuestas, el diseño del plan de muestreo tiene una importancia crítica en la determinación de qué datos ya existentes se recolectarán. Cuando se emplean experimentos, el diseño experimental tiene una importancia vital en la determinación de cuáles son los datos que serán generados o creados.

En el muestreo de encuestas pueden presentarse dos tipos de errores: el error muestral y los errores no muestrales. El error muestral es el tipo de error que se presenta debido a que para estimar un parámetro poblacional se emplea una muestra y no toda la población. Los errores no muestrales se refieren a *todos* los otros tipos de errores que pueden presentarse, tales como errores de medición, del entrevistador, de falta de respuesta y errores de procesamiento. Los errores no muestrales se controlan a través del diseño del cuestionario, de la capacitación de los entrevistadores, de la verificación cuidadosa de los datos, etc. El error muestral se minimiza mediante la elección adecuada del diseño y del tamaño de la muestra.

En este capítulo se habló de cuatro diseños muestrales muy empleados: muestreo aleatorio simple, muestreo aleatorio simple estratificado, muestreo por conglomerados y muestreo sistemático. El objetivo del diseño muestral es obtener la estimación más precisa al mínimo costo. Cuando la población es dividible en estratos, de manera que los elementos dentro de cada estrato sean relativamente homogéneos, el muestreo aleatorio simple estratificado permitirá obtener mayor precisión (intervalos de confianza más estrechos) que el muestreo aleatorio simple. Cuando los elementos se

pueden agrupar en conglomerados de manera que todos los elementos de un conglomerado se encuentren cercanos unos de otros geográficamente, el muestreo por conglomerados suele reducir el costo del entrevistador; en estos casos, el muestreo por conglomerados proporcionará mayor precisión a menor costo. El muestreo aleatorio sistemático se presentó como alternativa al muestreo aleatorio simple.

Glosario

Elemento Entidad de la que se recolectan los datos.
Población El conjunto de todos los elementos de interés.
Muestra Un subconjunto de la población.
Población objetivo Población acerca de la cual se quieren hacer las inferencias.
Población muestreada Población de la que se toma la muestra.
Unidades muestrales Las unidades que se toman para la muestra.
Marco Lista de las unidades muestrales para un estudio. La muestra es aleatoria para seleccionar unidades del marco.
Muestreo probabilístico Todo método de muestreo en el que se puede calcular la probabilidad que tiene cada posible muestra.
Muestreo no probabilístico Todo método de muestreo en el que no se puede calcular la probabilidad de seleccionar una determinada muestra.
Muestreo de conveniencia Un método no probabilístico de muestreo en el que los elementos se toman con base en la conveniencia.
Muestreo subjetivo Un método no probabilístico de muestreo en el que los elementos se seleccionan con base en el criterio de la persona que diseña el estudio.
Error muestral El error que se presenta debido a que se emplea una muestra y no toda la población para estimar un parámetro poblacional.
Error no muestral Todos los tipos de errores que no son un error muestral, como errores de medición, errores del entrevistador y errores de procesamiento.
Muestra aleatoria simple Muestra que se toma de tal manera que toda muestra de tamaño n tiene la misma probabilidad de ser elegida.
Cota del error muestral Número que se suma o que se resta de una estimación puntual para obtener un intervalo de confianza de aproximadamente 95%. La cota del error muestral es igual al doble del error estándar del estimador puntual.
Muestreo aleatorio simple estratificado Método probabilístico para tomar una muestra en el que, primero, se divide la población en estratos y después de cada estrato se toma una muestra aleatoria simple.
Muestreo por conglomerados Método probabilístico de muestreo en el que primero se divide la población en conglomerados y después se selecciona uno o más de los conglomerados para la muestra. En el muestreo por conglomerados en una sola etapa, se toman en la muestra todos los elementos de cada uno de los conglomerados elegidos; en el muestreo por conglomerados en dos etapas se toma una muestra de los elementos de cada uno de los conglomerados elegidos.
Muestreo sistemático Método para tomar una muestra en el que el primer elemento se toma aleatoriamente y después se toma cada k-ésimo elemento.

Fórmulas clave

Muestreo aleatorio simple

Estimación por intervalo de la media poblacional

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \quad \textbf{(22.1)}$$

Estimación del error estándar de la media

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right) \tag{22.2}$$

Intervalo de estimación para la media poblacional

$$\bar{x} \pm t_{\alpha/2} s_{\bar{x}} \tag{22.3}$$

Intervalo de confianza de aproximadamente 95% para estimar la media poblacional

$$\bar{x} \pm 2 s_{\bar{x}} \tag{22.4}$$

Estimación puntual del total poblacional

$$\hat{X} = N\bar{x} \tag{22.5}$$

Estimación del error estándar de $\hat{X}$

$$s_{\hat{X}} = N s_{\bar{x}} \tag{22.6}$$

Intervalo de confianza de aproximadamente 95% para estimar el total poblacional

$$N\bar{x} \pm 2 s_{\hat{X}} \tag{22.8}$$

Estimación del error estándar de la proporción

$$s_{\bar{p}} = \sqrt{\left(\frac{N-n}{N}\right)\left(\frac{\bar{p}(1-\bar{p})}{n-1}\right)} \tag{22.9}$$

Intervalo de confianza de aproximadamente 95% para estimar la proporción poblacional

$$\bar{p} \pm 2 s_{\bar{p}} \tag{22.10}$$

Tamaño de la muestra en una estimación de la media poblacional

$$n = \frac{Ns^2}{N\left(\frac{B^2}{4}\right) + s^2} \tag{22.12}$$

Tamaño de la muestra en una estimación del total poblacional

$$n = \frac{Ns^2}{\left(\frac{B^2}{4N}\right) + s^2} \tag{22.13}$$

Tamaño de la muestra en una estimación de la proporción poblacional

$$n = \frac{N\bar{p}(1-\bar{p})}{N\left(\frac{B^2}{4}\right) + \bar{p}(1-\bar{p})} \tag{22.14}$$

Muestreo aleatorio simple estratificado

Estimación puntual de la media poblacional

$$\bar{x}_{st} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \bar{x}_h \tag{22.15}$$

Estimación del error estándar de la media

$$s_{\bar{x}_{st}} = \sqrt{\frac{1}{N^2} \sum_{h=1}^{H} N_h(N_h - n_h)\frac{s_h^2}{n_h}} \tag{22.16}$$

Intervalo de confianza de aproximadamente 95% para la media poblacional

$$\bar{x}_{st} \pm 2s_{\bar{x}_{st}} \tag{22.17}$$

Estimación puntual del total poblacional

$$\hat{X} = N\bar{x}_{st} \tag{22.18}$$

Estimación del error estándar de $\hat{X}$

$$s_{\hat{X}} = Ns_{\bar{x}_{st}} \tag{22.19}$$

Intervalo de confianza de aproximadamente 95% para estimar el total poblacional

$$N\bar{x}_{st} \pm 2s_{\hat{X}} \tag{22.20}$$

Estimación puntual de la proporción poblacional

$$\bar{p}_{st} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \bar{p}_h \tag{22.21}$$

Estimación del error estándar de $\bar{p}_{st}$

$$s_{\bar{p}_{st}} = \sqrt{\frac{1}{N^2} \sum_{h=1}^{H} N_h(N_h - n_h)\left[\frac{\bar{p}_h(1 - \bar{p}_h)}{n_h - 1}\right]} \tag{22.22}$$

Intervalo de confianza de aproximadamente 95% para la proporción poblacional

$$\bar{p}_{st} \pm 2s_{\bar{p}_{st}} \tag{22.23}$$

Asignación del total n de la muestra a los estratos: asignación de Neyman

$$n_h = n\left(\frac{N_h s_h}{\sum_{h=1}^{H} N_h s_h}\right) \tag{22.24}$$

Tamaño de la muestra cuando se estima la media poblacional

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{N^2\left(\frac{B^2}{4}\right) + \sum_{h=1}^{H} N_h s_h^2} \tag{22.25}$$

Tamaño de la muestra cuando se estima el total poblacional

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{\frac{B^2}{4} + \sum_{h=1}^{H} N_h s_h^2} \qquad \textbf{(22.26)}$$

Tamaño de la muestra para la estimación de la proporción poblacional

$$n = \frac{\left(\sum_{h=1}^{H} N_h \sqrt{\bar{p}_h(1 - \bar{p}_h)}\right)^2}{N^2\left(\frac{B^2}{4}\right) + \sum_{h=1}^{H} N_h \bar{p}_h(1 - \bar{p}_h)} \qquad \textbf{(22.27)}$$

Asignación proporcional de la muestra *n* a los estratos

$$n_h = n\left(\frac{N_h}{N}\right) \qquad \textbf{(22.28)}$$

Muestreo por conglomerados

Estimación puntual de la media poblacional

$$\bar{x}_c = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i} \qquad \textbf{(22.29)}$$

Estimación del error estándar de la media

$$s_{\bar{x}_c} = \sqrt{\left(\frac{N - n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(x_i - \bar{x}_c M_i)^2}{n - 1}} \qquad \textbf{(22.30)}$$

Intervalo de confianza de aproximadamente 95% para la media poblacional

$$\bar{x}_c \pm 2s_{\bar{x}_c} \qquad \textbf{(22.31)}$$

Estimación puntual del total poblacional

$$\hat{X} = M\bar{x}_c \qquad \textbf{(22.32)}$$

Estimación del error estándar de $\hat{X}$

$$s_{\hat{X}} = Ms_{\bar{x}_c} \qquad \textbf{(22.33)}$$

Intervalo de confianza de aproximadamente 95% para el total poblacional

$$M\bar{x}_c \pm 2s_{\hat{X}} \qquad \textbf{(22.34)}$$

Estimación puntual de la proporción poblacional

$$\bar{p}_c = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i} \qquad \textbf{(22.35)}$$

Estimación del error estándar de $\bar{p}_c$

$$s_{\bar{p}_c} = \sqrt{\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(a_i - \bar{p}_c M_i)^2}{n-1}} \tag{22.36}$$

Intervalo de confianza de aproximadamente 95% para estimar la proporción poblacional

$$\bar{p}_c \pm 2 s_{\bar{p}_c} \tag{22.37}$$

Ejercicios complementarios

18. Para evaluar la aceptación que tiene ante los consumidores una nueva publicidad de la cerveza Millar Lite Beer, Louis Harris realizó una encuesta nacional con 363 adultos que habían visto la nueva publicidad de Millar Lite (*USA Today,* 17 de noviembre de 1997). Algunas de las respuestas obtenidas en la encuesta fueron las que se presentan a continuación (*Nota*: Como en esta encuesta sólo se tomó en la muestra una pequeña fracción de todos los adultos, en todas las fórmulas en las que se use el error estándar suponga que $((N - n)/N = 1)$.
    a. Diecinueve por ciento de los entrevistados indicó que la nueva publicidad les gustaba mucho. Dé un intervalo de confianza de 95% para esta proporción poblacional.
    b. A 31% de los entrevistados no les gustó la nueva publicidad. Dé un intervalo de confianza de 95% para esta proporción poblacional.
    c. Diecisiete por ciento de los entrevistados indicó que la nueva publicidad les parecía muy efectiva. Dé un intervalo de confianza de 95% para la proporción de adultos que encuentra que la nueva publicidad es muy efectiva.
    d. Louis Harris informó que "el margen de error era de cinco puntos porcentuales". ¿Qué significa esto y cómo cree usted que llegó a esta cifra?
    e. ¿Cómo pueden sesgar los errores no muestrales los resultados de un estudio de este tipo?
19. Mediante una encuesta entre los suscriptores de su edición interactiva, *The Wall Street Journal* realizó una investigación. Una de las preguntas que se hizo a los 504 encuestados era si usaban su laptop cuando viajaban; 55% respondió que sí. Otra de las preguntas era si cuando viajaban empleaban un servicio exprés o un servicio de paquetería; 31% respondió que sí (*The Wall Street Journal Interactive Edition Subscriber Survey*, 2000).
    a. Estime el error estándar para la proporción que usa la laptop.
    b. Estime el error estándar de la proporción que usa un servicio exprés o un servicio de paquetería.
    c. ¿Son iguales las estimaciones del error estándar dadas en el inciso a y en el inciso b? Si son diferentes explique por qué.
    d. Dé un intervalo de confianza de 95% para la proporción que usa la laptop.
    e. Dé un intervalo de confianza de 95% para la proporción que usa un servicio exprés o un servicio de paquetería.
20. Mediante una encuesta se realizó un estudio sobre la calidad de vida de los empleados de una fábrica. A 300 de los 3 000 empleados de la fábrica se les envió un cuestionario. La tasa de respuesta fue de 67%, lo que corresponde a 200 cuestionarios contestados.
    a. En la muestra el salario anual medio fue $\bar{x} = \$23\ 200$ con $s = \$3\ 000$. Dé un intervalo de confianza de aproximadamente 95% para el salario anual medio de la población.
    b. Use la información del inciso a para obtener un intervalo de confianza de aproximadamente 95% para el total de los salarios de los 3 000 empleados.

c. Setenta y tres por ciento de los entrevistados informó estar "en general satisfechos" con su trabajo. Proporcione un intervalo de confianza de aproximadamente 95% para esta proporción poblacional.

d. Haga un comentario sobre si usted considera que el resultado del inciso c pueda estar sesgado. ¿Cambiaría su opinión si supiera que a los entrevistados se les garantizó el anonimato?

21. En un informe del Comité Judicial del Senado de Estados Unidos se presenta el número de homicidios en cada estado. En Indiana, Ohio y Kentucky, el número de homicidios fue, respectivamente, 380, 760 y 260. Suponga que se tomó una muestra aleatoria estratificada para conocer más acerca de las víctimas y de la causa de su muerte, los resultados se presentan a continuación

Estrato	Tamaño de la muestra	Disparo	Golpiza	Víctima urbana
Indiana	30	10	9	21
Ohio	45	19	12	34
Kentucky	25	7	11	15

a. Dé un intervalo de confianza de aproximadamente 95% para la proporción de muertes por disparo con arma de fuego en Indiana.

b. Estime el número total de muertes por disparo con arma de fuego en Ohio.

c. Dé un intervalo de confianza de aproximadamente 95% para la proporción de muertes por disparo con arma de fuego en Ohio.

d. Dé un intervalo de confianza de aproximadamente 95% para la proporción de muertes por disparo con arma de fuego en los tres estados.

22. Remítase a los datos del ejercicio 21.

a. Estime la cantidad total de muertes (en los tres estados) por golpizas.

b. Dé un intervalo de confianza de aproximadamente 95% para la proporción de muertes por golpizas en los tres estados.

c. Dé un intervalo de confianza de aproximadamente 95% para la proporción de víctimas urbanas.

d. Estime la cantidad total de víctimas urbanas.

23. Se va a tomar una muestra aleatoria simple estratificada de los clientes de un banco para tener información sobre actitudes y datos demográficos. La estratificación se basará en el estado de cuenta al 30 de junio de 2001. A continuación se presenta una distribución de frecuencias para cada estrato junto con las desviaciones estándar de los estados de cuenta por estrato.

Estrato ($)	Cuentas	Desviación estándar de los estados de cuenta
0.00–1 000.00	3000	80
1 000.01–2 000.00	600	150
2 000.01–5 000.00	250	220
5 000.01–10 000.00	100	700
más de 10 000.00	50	3000

a. Si el costo por unidad muestreada es aproximadamente el mismo en todos los estratos, determine el número total de personas que deberán incluirse en la muestra. Suponga que se desea una cota del error de estimación de la media poblacional de los estados de cuenta de $B = \$20$.

b. Utilice el procedimiento de asignación de Neyman para determinar el número que debe ser muestreado de cada estrato.

24. Un organismo público está interesado en conocer más acerca de las personas que viven en casas de reposo en una determinada ciudad. En esa ciudad hay en total 100 casas de reposo que atienden a 4 800 personas y se ha tomado una muestra por conglomerados de seis casas de reposo.

Casa	Residentes	Edad promedio de los residentes	Residentes inválidos
1	14	61	12
2	7	74	2
3	96	78	30
4	23	69	8
5	71	73	10
6	29	84	22

a. Estime la edad promedio de los residentes en las casas de reposo en esa ciudad.
b. Dé un intervalo de confianza de aproximadamente 95% para la proporción de personas inválidas en las casas de reposo de esa ciudad.
c. Estime el número total de personas inválidas en las casas de reposo de esa ciudad.



Apéndices



Apéndice A: Referencias y bibliografía

General

Bowerman, B. L. y O'Connell, R. T., *Applied Statistics: Improving Business Processes*, Irwin, 1996.
Freedman, D., Pisani, R. y Purves, R., *Statistics*, 3a. ed., W. W. Norton, 1997.
Hogg, R. V. y Craig, A. T., *Introduction to Mathematical Statistics*, 5a. ed., Prentice Hall, 1994.
Hogg, R. V. y Tanis, E. A., *Probability and Statistical Inference*, 6a. ed., Prentice Hall, 2001.
Joiner, B. L. y Ryan, B. F., *Minitab Handbook*, Brooks/Cole. 2000.
Miller, I. y Miller, M. *John E. Freund's Mathematical Statistics*, Prentice Hall, 1998.
Moore, D. S. y McCabe, G. P., *Introduction to the Practice of Statistics*, 4a. ed., Freeman, 2003.
Roberts, H., *Data Analysis for Managers with Minitab*, Scientific Press, 1991.
Tanur, J. M. *Statistics: A Guide to the Unknown*, 4a. ed. Brooks-/Cole, 2002.
Tukey, J. W., *Exploratory Data Analysis*, Addison-Wesley, 1977.

Diseño experimental

Cochran, W. G. y Cox, G. M., *Experimental Designs*, 2a. ed., Wiley, 1992.
Hicks, C. R. y Turner, K. V. *Fundamental Concepts in the Design of Experiments*, 5a. ed., Oxford University Press, 1999.
Montgomery, D. C., *Design and Analysis of Experiments*, 5a. ed., Wiley, 2000.
Winer, B. J., Michels, K. M. y Brown, D. R., *Statistical Principles in Experimental Design*, 3a. ed., McGraw-Hill, 1991.
Wu, C. F. y Hamada, M., *Experiments: Planning Analysis and Parameter Optimization*, Wiley, 2000.

Pronóstico

Bowerman, B. L. y O'Connell, R. T., *Forecasting and Time Series: An Applied Approach*, 3a. ed., Brooks/Cole, 2000.
Box, G. E. P., Reinsel, G. C. y Jenkins, G., *Time Series Analysis: Forecasting and Control*, 3a. ed., Prentice Hall, 1994.
Makridakis, S., Wheelwright, S. C. y Hyndman, R. J., *Forecasting: Methods and Applications*, 3a. ed., Wiley, 1997.

Números índice

U.S. Department of Commerce. *Survey of Current Business.*
U.S. Department of Labor, Bureau of Labor Statistics. *CPI Detailed Report.*
U.S. Department of Labor. *Producer Price Indexes.*

Métodos no paramétricos

Conover, W: J., *Practical Nonparametric Statistics,* 3a. ed., Wiley, 1998.
Gibbons, J. D. y Chakraborti, S., *Nonparametric Statistical Inference*, 3a. ed., Marcel Dekker, 1992.
Siegel, S. y Castellan, N. J., *Nonparametric Statistics for the Behavioral Sciences,* 2a. ed., McGraw-Hill, 1990.
Sprent, P., *Applied Non-Parametric Statistical Methods*, CRC, 1993.

Probabilidad

Hogg, R. V. y Tanis, E. A., *Probability and Statistical Inference*, 6a. ed., Prentice Hall, 2001.
Ross, S. M., *Introduction to Probability Models*, 7a. ed., Academic Press, 2000.
Wackerly, D. D., Mendelhall, W. y Scheaffer, R. L., *Mathematical Statistics with Applications,* 6a. ed., Duxbury Press, 2002.

Control de calidad

Deming, W. E. *Quality, Productivity, and Competitive Position*, MIT, 1982.
Evans, J. R. y Lindsay, W. M., *The Management and Control Quality*, 5a. ed., South-Western, 2001.
Gryna F. M. y Juran, I. M., *Quality Planning and Analysis: From Product Development Through Use*, 3a. ed., McGraw-Hill, 1993.
Ishikawa, K., *Introduction to Quality Control*, Kluwer Academic, 1991.
Montgomery, D. C., *Introduction to Statistical Quality Control*, 4a. ed., Wiley, 2000.

Análisis de regresión

Belsley, D. A., *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, Wiley, 1991.
Chatterjee, S. y Price, B., *Regression Analysis by Example*, 3a. ed., Wiley, 1999.
Draper, N. R. y Smith, H., *Applied Regression Analysis*, 3a. ed., Wiley, 1998.
Graybill, F. A. y Iyer, H., *Regression Analysis: Concepts and Applications*, Duxbury Press, 1994.
Hosmer, D. W. y Lemeshow, S., *Applies Logistic Regression*, 2a. ed., Wiley, 2000.
Kleinbaum, D. G., Kupper, L. L. y Muller, K. E., *Applied Regression Analysis and Other Multivariate Methods*, 3a. ed., Duxbury Press, 1997.

Kutner, M. H., Nachtschiem, C. J., Wasserman, W. y Neter, J., *Applied Linear Statistical Models*, 4a. ed., Irwin, 1996.

Mendenhall, M. y Sincich, T., *A Second Course in Statistics: Regression Analysis*, 5a. ed., Prentice Hall, 1996.

Myers, R. H., *Classical and Modern Regression with Applications,* 2a. ed., PWS, 1990.

Análisis de decisión

Chernoff, H. y Moses, L. E., *Elementary Decision Theory*, Dover, 1987.

Clemen, R. T. y Reilly, T., *Making Hard Decisions with Decision Tools*, Duxbury Press, 2001.

Goodwin, P. y Wright, G., *Decision Analysis for Management Judgment*, 2a. ed., Wiley, 1999.

Pratt, J. W., Raiffa, H. y Schlaifer, R., *Introduction to Statistical Decision Theory*, MIT Press, 1995.

Raiffa, H., *Decision Analysis*, McGraw-Hill, 1997

Muestreo

Cochran, W. G., *Sampling Techniques*, 3a. ed., Wiley, 1977.

Deming, W. E., *Some Theory of Sampling*, Dover, 1984.

Hansen, M. H., Hurwitz, W. N., Madow, W. G. y Hanson, M. N., *Simple Survey Methods and Theory*, Wiley, 1993.

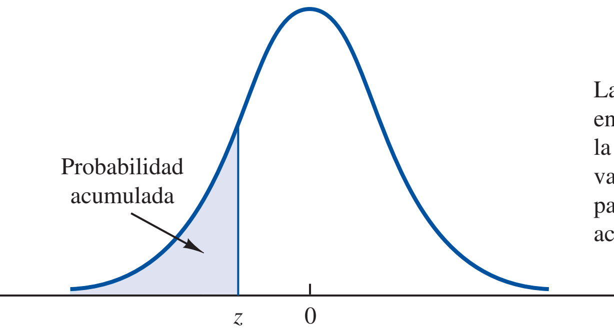
Kish, L., *Survey Sampling*, Wiley, 1995.

Levy, P. S. y Lemeshow, S., *Sampling of Populations: Methods and Applications*, 3a. ed., Wiley, 1999.

Scheaffer, R. L., Mendenhall, W. y Ott, L., *Elementary Survey Sampling,* 5a. ed., Duxbury Press, 1996.

Apéndice B: Tablas

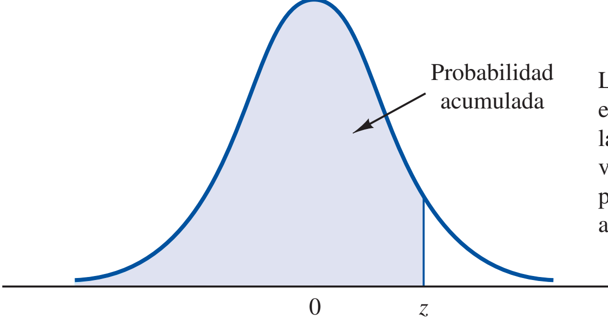
TABLA 1 PROBABILIDADES ACUMULADAS EN LA DISTRIBUCIÓN NORMAL ESTÁNDAR



Las entradas que aparecen en la tabla dan el área bajo la curva y a la izquierda del valor de z. Por ejemplo, para $z = -0.85$, la probabilidad acumulada es 0.1977.

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

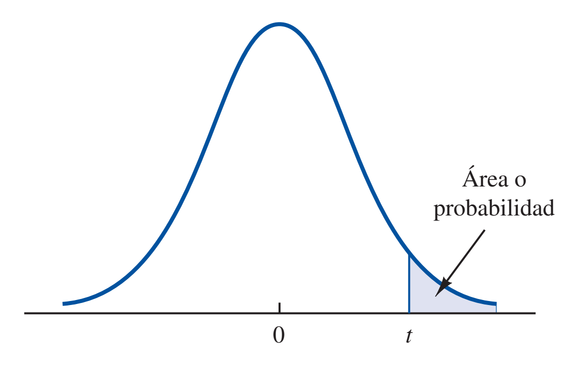
TABLA 1 PROBABILIDADES ACUMULADAS EN LA DISTRIBUCIÓN NORMAL ESTÁNDAR (*continuación*)



Las entradas que aparecen en la tabla dan el área bajo la curva y a la izquierda del valor de z. Por ejemplo, para $z = 1.25$, la probabilidad acumulada es 0.8944.

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9913
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9986	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

TABLA 2 DISTRIBUCIÓN *t*



Las entradas que aparecen en la tabla dan los valores de t correspondientes a un área o probabilidad en la cola superior de la distribución t. Por ejemplo, para 10 grados de libertad y un área de 0.05 en la cola superior, $t_{0.05} = 1.812$.

Grados de libertad	Área en la cola superior					
	0.20	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	0.978	1.638	2.353	3.182	4.541	5.841
4	0.941	1.533	2.132	2.776	3.747	4.604
5	0.920	1.476	2.015	2.571	3.365	4.032
6	0.906	1.440	1.943	2.447	3.143	3.707
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1.397	1.860	2.306	2.896	3.355
9	0.883	1.383	1.833	2.262	2.821	3.250
10	0.879	1.372	1.812	2.228	2.764	3.169
11	0.876	1.363	1.796	2.201	2.718	3.106
12	0.873	1.356	1.782	2.179	2.681	3.055
13	0.870	1.350	1.771	2.160	2.650	3.012
14	0.868	1.345	1.761	2.145	2.624	2.977
15	0.866	1.341	1.753	2.131	2.602	2.947
16	0.865	1.337	1.746	2.120	2.583	2.921
17	0.863	1.333	1.740	2.110	2.567	2.898
18	0.862	1.330	1.734	2.101	2.552	2.878
19	0.861	1.328	1.729	2.093	2.539	2.861
20	0.860	1.325	1.725	2.086	2.528	2.845
21	0.859	1.323	1.721	2.080	2.518	2.831
22	0.858	1.321	1.717	2.074	2.508	2.819
23	0.858	1.319	1.714	2.069	2.500	2.807
24	0.857	1.318	1.711	2.064	2.492	2.797
25	0.856	1.316	1.708	2.060	2.485	2.787
26	0.856	1.315	1.706	2.056	2.479	2.779
27	0.855	1.314	1.703	2.052	2.473	2.771
28	0.855	1.313	1.701	2.048	2.467	2.763
29	0.854	1.311	1.699	2.045	2.462	2.756
30	0.854	1.310	1.697	2.042	2.457	2.750
31	0.853	1.309	1.696	2.040	2.453	2.744
32	0.853	1.309	1.694	2.037	2.449	2.738
33	0.853	1.308	1.692	2.035	2.445	2.733
34	0.852	1.307	1.691	2.032	2.441	2.728

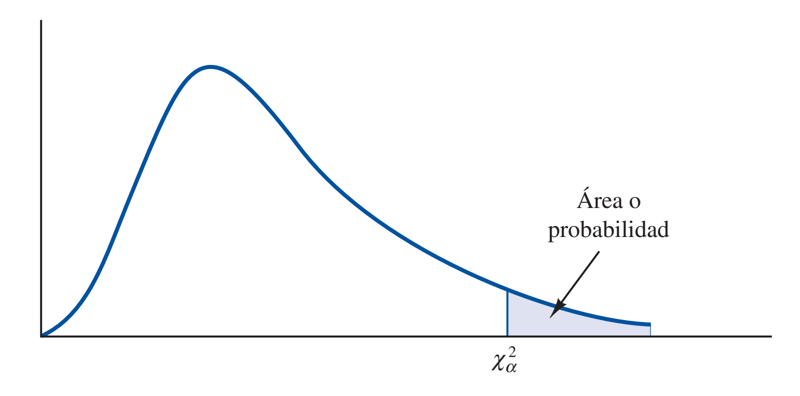
TABLA 2 DISTRIBUCIÓN *t* (*continuación*)

Grados de libertad	Área en la cola superior					
	0.20	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
35	0.852	1.306	1.690	2.030	2.438	2.724
36	0.852	1.306	1.688	2.028	2.434	2.719
37	0.851	1.305	1.687	2.026	2.431	2.715
38	0.851	1.304	1.686	2.024	2.429	2.712
39	0.851	1.304	1.685	2.023	2.426	2.708
40	0.851	1.303	1.684	2.021	2.423	2.704
41	0.850	1.303	1.683	2.020	2.421	2.701
42	0.850	1.302	1.682	2.018	2.418	2.698
43	0.850	1.302	1.681	2.017	2.416	2.695
44	0.850	1.301	1.680	2.015	2.414	2.692
45	0.850	1.301	1.679	2.014	2.412	2.690
46	0.850	1.300	1.679	2.013	2.410	2.687
47	0.849	1.300	1.678	2.012	2.408	2.685
48	0.849	1.299	1.677	2.011	2.407	2.682
49	0.849	1.299	1.677	2.010	2.405	2.680
50	0.849	1.299	1.676	2.009	2.403	2.678
51	0.849	1.298	1.675	2.008	2.402	2.676
52	0.849	1.298	1.675	2.007	2.400	2.674
53	0.848	1.298	1.674	2.006	2.399	2.672
54	0.848	1.297	1.674	2.005	2.397	2.670
55	0.848	1.297	1.673	2.004	2.396	2.668
56	0.848	1.297	1.673	2.003	2.395	2.667
57	0.848	1.297	1.672	2.002	2.394	2.665
58	0.848	1.296	1.672	2.002	2.392	2.663
59	0.848	1.296	1.671	2.001	2.391	2.662
60	0.848	1.296	1.671	2.000	2.390	2.660
61	0.848	1.296	1.670	2.000	2.389	2.659
62	0.847	1.295	1.670	1.999	2.388	2.657
63	0.847	1.295	1.669	1.998	2.387	2.656
64	0.847	1.295	1.669	1.998	2.386	2.655
65	0.847	1.295	1.669	1.997	2.385	2.654
66	0.847	1.295	1.668	1.997	2.384	2.652
67	0.847	1.294	1.668	1.996	2.383	2.651
68	0.847	1.294	1.668	1.995	2.382	2.650
69	0.847	1.294	1.667	1.995	2.382	2.649
70	0.847	1.294	1.667	1.994	2.381	2.648
71	0.847	1.294	1.667	1.994	2.380	2.647
72	0.847	1.293	1.666	1.993	2.379	2.646
73	0.847	1.293	1.666	1.993	2.379	2.645
74	0.847	1.293	1.666	1.993	2.378	2.644
75	0.846	1.293	1.665	1.992	2.377	2.643
76	0.846	1.293	1.665	1.992	2.376	2.642
77	0.846	1.293	1.665	1.991	2.376	2.641
78	0.846	1.292	1.665	1.991	2.375	2.640
79	0.846	1.292	1.664	1.990	2.374	2.639

TABLA 2 DISTRIBUCIÓN *t* (*continuación*)

Grados de libertad	Área en la cola superior					
	0.20	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
80	0.846	1.292	1.664	1.990	2.374	2.639
81	0.846	1.292	1.664	1.990	2.373	2.638
82	0.846	1.292	1.664	1.989	2.373	2.637
83	0.846	1.292	1.663	1.989	2.372	2.636
84	0.846	1.292	1.663	1.989	2.372	2.636
85	0.846	1.292	1.663	1.988	2.371	2.635
86	0.846	1.291	1.663	1.988	2.370	2.634
87	0.846	1.291	1.663	1.988	2.370	2.634
88	0.846	1.291	1.662	1.987	2.369	2.633
89	0.846	1.291	1.662	1.987	2.369	2.632
90	0.846	1.291	1.662	1.987	2.368	2.632
91	0.846	1.291	1.662	1.986	2.368	2.631
92	0.846	1.291	1.662	1.986	2.368	2.630
93	0.846	1.291	1.661	1.986	2.367	2.630
94	0.845	1.291	1.661	1.986	2.367	2.629
95	0.845	1.291	1.661	1.985	2.366	2.629
96	0.845	1.290	1.661	1.985	2.366	2.628
97	0.845	1.290	1.661	1.985	2.365	2.627
98	0.845	1.290	1.661	1.984	2.365	2.627
99	0.845	1.290	1.660	1.984	2.364	2.626
100	0.845	1.290	1.660	1.984	2.364	2.626
∞	0.842	1.282	1.645	1.960	2.326	2.576

TABLA 3 DISTRIBUCIÓN CHI-CUADRADA



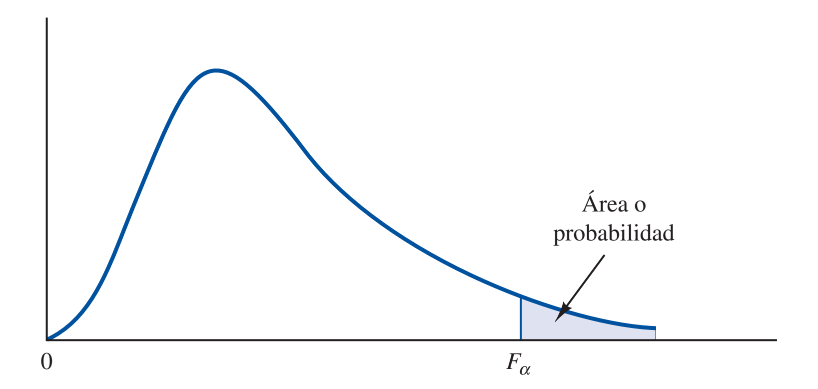
Las entradas que aparecen en la tabla dan los valores de χ^2_α, donde α es el área o probabilidad en la cola superior de la distribución chi-cuadrada. Por ejemplo, para 10 grados de libertad y un área de 0.01 en la cola superior, $\chi_{0.01}{}^2 =$ 23.209.

Grados de libertad	Áreas en la cola superior									
	0.995	**0.99**	**0.975**	**0.95**	**0.90**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.994
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.335

TABLA 3 DISTRIBUCIÓN CHI-CUADRADA (*continuación*)

Grados de libertad	Áreas en la cola superior									
	0.995	**0.99**	**0.975**	**0.95**	**0.90**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
45	24.311	25.901	28.366	30.612	33.350	57.505	61.656	65.410	69.957	73.166
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
55	31.735	33.571	36.398	38.958	42.060	68.796	73.311	77.380	82.292	85.749
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
65	39.383	41.444	44.603	47.450	50.883	79.973	84.821	89.177	94.422	98.105
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
75	47.206	49.475	52.942	56.054	59.795	91.061	96.217	100.839	106.393	110.285
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
85	55.170	57.634	61.389	64.749	68.777	102.079	107.522	112.393	118.236	122.324
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
95	63.250	65.898	69.925	73.520	77.818	113.038	118.752	123.858	129.973	134.247
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.170

TABLA 4 DISTRIBUCIÓN F



Las entradas que aparecen en la tabla dan los valores de F_α, donde α es el área o probabilidad en la cola superior de la distribución F. Por ejemplo, para 4 grados de libertad en el numerador, 8 grados de libertad en el denominador y un área de 0.05 en la cola superior, $F_{0.05} = 3.84$.

Grados de libertad en el denominador	Área en la cola superior	Grados de libertad en el numerador																	
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	100	1000
1	0.10	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	61.22	61.74	62.05	62.26	62.53	62.79	63.01	63.30
	0.05	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95	248.02	249.26	250.10	251.14	252.20	253.04	254.19
	0.025	647.79	799.48	864.15	899.60	921.83	937.11	948.20	956.64	963.28	968.63	984.87	993.08	998.09	1001.40	1005.60	1009.79	1013.16	1017.76
	0.01	4052.18	4999.34	5403.53	5624.26	5763.96	5858.95	5928.33	5980.95	6022.40	6055.93	6156.97	6208.66	6239.86	6260.35	6286.43	6312.97	6333.92	6362.80
2	0.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.46	19.47	19.48	19.49	19.49
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
	0.01	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.43	99.45	99.46	99.47	99.48	99.48	99.49	99.50
3	0.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20	5.18	5.17	5.17	5.16	5.15	5.14	5.13
	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.57	8.55	8.53
	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25	14.17	14.12	14.08	14.04	13.99	13.96	13.91
	0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	26.87	26.69	26.58	26.50	26.41	26.32	26.24	26.14
4	0.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.56	8.50	8.46	8.41	8.36	8.32	8.26
	0.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20	14.02	13.91	13.84	13.75	13.65	13.58	13.47
5	0.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.324	3.21	3.19	3.17	3.16	3.14	3.13	3.11
	0.05	6.61.	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.43	4.41	4.37
	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.33	6.27	6.23	6.18	6.12	6.08	6.02
	0.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72	9.55	9.45	9.38	9.29	9.20	9.13	9.03

TABLA 4 DISTRIBUCIÓN *F* (*continuación*)

Grados de libertad en el denominador	Área en la cola superior	Grados de libertad en el numerador																	
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**15**	**20**	**25**	**30**	**40**	**60**	**100**	**1000**
6	0.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87	2.84	2.81	2.80	2.78	2.76	2.75	2.72
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.74	3.71	3.67
	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.17	5.11	5.07	5.01	4.96	4.92	4.86
	0.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.30	7.23	7.14	7.06	6.99	6.89
7	0.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63	2.59	2.57	2.56	2.54	2.51	2.50	2.47
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.30	3.27	3.23
	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.47	4.40	4.36	4.31	4.25	4.21	4.15
	0.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	6.06	5.99	5.91	5.82	5.75	5.66
8	0.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.30
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.01	2.97	2.93
	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	4.00	3.94	3.89	3.84	3.78	3.74	3.68
	0.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.26	5.20	5.12	5.03	4.96	4.87
9	0.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34	2.30	2.27	2.25	2.23	2.21	2.19	2.16
	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.79	2.76	2.71
	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.67	3.60	3.56	3.51	3.45	3.40	3.34
	0.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.71	4.65	4.57	4.48	4.41	4.32
10	0.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24	2.20	2.17	2.16	2.13	2.11	2.09	2.06
	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.62	2.59	2.54
	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.42	3.35	3.31	3.26	3.20	3.15	3.09
	0.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.31	4.25	4.17	4.08	4.01	3.92
11	0.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.17	2.12	2.10	2.08	2.05	2.03	2.01	1.98
	0.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.49	2.46	2.41
	0.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33	3.23	3.16	3.12	3.06	3.00	2.96	2.89
	0.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	4.01	3.94	3.86	3.78	3.71	3.61
12	0.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.10	2.06	2.03	2.01	1.99	1.96	1.94	1.91
	0.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.38	2.35	2.30
	0.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18	3.07	3.01	2.96	2.91	2.85	2.80	2.73
	0.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.76	3.70	3.62	3.54	3.47	3.37
13	0.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
	0.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.30	2.26	2.21
	0.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05	2.95	2.88	2.84	2.78	2.72	2.67	2.60
	0.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.57	3.51	3.43	3.34	3.27	3.18
14	0.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.01	1.96	1.93	1.99	1.89	1.86	1.83	1.80
	0.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.22	2.19	2.14
	0.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95	2.84	2.78	2.73	2.67	2.61	2.56	2.50
	0.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.41	3.35	3.27	3.18	3.11	3.02
15	0.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97	1.92	1.89	1.87	1.85	1.82	1.79	1.76
	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.16	2.12	2.07
	0.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.76	2.69	2.64	2.59	2.52	2.47	2.40
	0.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.28	3.21	3.13	3.05	2.98	2.88

Grados de libertad en el denominador	Área en la cola superior	Grados de libertad en el numerador																	
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	100	1000
16	0.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.94	1.89	1.86	1.84	1.81	1.78	1.76	1.72
	0.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.23	2.19	2.15	2.11	2.07	2.02
	0.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.79	2.68	2.61	2.57	2.51	2.45	2.40	2.32
	0.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.26	3.16	3.10	3.02	2.93	2.86	2.76
17	0.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.91	1.86	1.83	1.81	1.78	1.75	1.73	1.69
	0.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.18	2.15	2.10	2.06	2.02	1.97
	0.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.72	2.62	2.55	2.50	2.44	2.38	2.33	2.26
	0.01	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31	3.16	3.07	3.00	2.92	2.83	2.76	2.66
18	0.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.89	1.84	1.80	1.78	1.75	1.72	1.70	1.66
	0.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.14	2.11	2.06	2.02	1.98	1.92
	0.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.67	2.56	2.49	2.44	2.38	2.32	2.27	2.20
	0.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	3.08	2.98	2.92	2.84	2.75	2.68	2.58
19	0.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.86	1.81	1.78	1.76	1.73	1.70	1.67	1.64
	0.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.11	2.07	2.03	1.98	1.94	1.88
	0.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.62	2.51	2.44	2.39	2.33	2.27	2.22	2.14
	0.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15	3.00	2.91	2.84	2.76	2.67	2.60	2.50
20	0.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.84	1.79	1.76	1.74	1.71	1.68	1.65	1.61
	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.95	1.91	1.85
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.46	2.40	2.35	2.29	2.22	2.17	2.09
	0.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.84	2.78	2.69	2.61	2.54	2.43
21	0.10	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.83	1.78	1.74	1.72	1.69	1.66	1.63	1.59
	0.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.18	2.10	2.05	2.01	1.96	1.92	1.88	1.82
	0.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.53	2.42	2.36	2.31	2.25	2.18	2.13	2.05
	0.01	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.03	2.88	2.79	2.72	2.64	2.55	2.48	2.37
22	0.10	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.81	1.76	1.73	1.70	1.67	1.64	1.61	1.57
	0.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	2.02	1.98	1.94	1.89	1.85	1.79
	0.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.50	2.39	2.32	2.27	2.21	2.14	2.09	2.01
	0.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98	2.83	2.73	2.67	2.58	2.50	2.42	2.32
23	0.10	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.80	1.74	1.71	1.69	1.66	1.62	1.59	1.55
	0.05	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.13	2.05	2.00	1.96	1.91	1.86	1.82	1.76
	0.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.47	2.36	2.29	2.24	2.18	2.11	2.06	1.98
	0.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	2.93	2.78	2.69	2.62	2.54	2.45	2.37	2.27
24	0.10	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.78	1.73	1.70	1.67	1.64	1.61	1.58	1.54
	0.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.97	1.94	1.89	1.84	1.80	1.74
	0.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.44	2.33	2.26	2.21	2.15	2.08	2.02	1.94
	0.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	2.89	2.74	2.64	2.58	2.49	2.40	2.33	2.22

TABLA 4 DISTRIBUCIÓN F (*continuación*)

Grados de libertad en el denominador	Área en la cola superior	Grados de libertad en el numerador																	
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**15**	**20**	**25**	**30**	**40**	**60**	**100**	**1000**
25	0.10	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.77	1.72	1.68	1.66	1.63	1.59	1.56	1.52
	0.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.09	2.01	1.96	1.92	1.87	1.82	1.78	1.72
	0.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.41	2.30	2.23	2.18	2.12	2.05	2.00	1.91
	0.01	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.85	2.70	2.60	2.54	2.45	2.36	2.29	2.18
26	0.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.76	1.71	1.67	1.65	1.61	1.58	1.55	1.51
	0.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.94	1.90	1.85	1.80	1.76	1.70
	0.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.39	2.28	2.21	2.16	2.09	2.03	1.97	1.89
	0.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.81	2.66	2.57	2.50	2.42	2.33	2.25	2.14
27	0.10	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.75	1.70	1.66	1.64	1.60	1.57	1.54	1.50
	0.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.06	1.97	1.92	1.88	1.84	1.79	1.74	1.68
	0.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.36	2.25	2.18	2.13	2.07	2.00	1.94	1.86
	0.01	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.78	2.63	2.54	2.47	2.38	2.29	2.22	2.11
28	0.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.74	1.69	1.65	1.63	1.59	1.56	1.53	1.48
	0.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.91	1.87	1.82	1.77	1.73	1.66
	0.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.34	2.23	2.16	2.11	2.05	1.98	1.92	1.84
	0.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.75	2.60	2.51	2.44	2.35	2.26	2.19	2.08
29	0.10	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.73	1.68	1.64	1.62	1.58	1.55	1.52	1.47
	0.05	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.03	1.94	1.89	1.85	1.81	1.75	1.71	1.65
	0.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.32	2.21	2.14	2.09	2.03	1.96	1.90	1.82
	0.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.73	2.57	2.48	2.41	2.33	2.23	2.16	2.05
30	0.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.72	1.67	1.63	1.61	1.57	1.54	1.51	1.46
	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.88	1.84	1.79	1.74	1.70	1.63
	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.20	2.12	2.07	2.01	1.94	1.88	1.80
	0.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.55	2.45	2.39	2.30	2.21	2.13	2.02
40	0.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.66	1.61	1.57	1.54	1.51	1.47	1.43	1.38
	0.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.78	1.74	1.69	1.64	1.59	1.52
	0.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.18	2.07	1.99	1.94	1.88	1.80	1.74	1.65
	0.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.37	2.27	2.20	2.11	2.02	1.94	1.82
60	0.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.60	1.54	1.50	1.48	1.44	1.40	1.36	1.30
	0.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.69	1.65	1.59	1.53	1.48	1.40
	0.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.06	1.94	1.87	1.82	1.74	1.67	1.60	1.49
	0.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.20	2.10	2.03	1.94	1.84	1.75	1.62
100	0.10	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.56	1.49	1.45	1.42	1.38	1.34	1.29	1.22
	0.05	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.68	1.62	1.57	1.52	1.45	1.39	1.30
	0.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97	1.85	1.77	1.71	1.64	1.56	1.48	1.36
	0.01	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22	2.07	1.97	1.89	1.80	1.69	1.60	1.45
1000	0.10	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64	1.61	1.49	1.43	1.38	1.35	1.30	1.25	1.20	1.08
	0.05	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.68	1.58	1.52	1.47	1.41	1.33	1.26	1.11
	0.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13	2.06	1.85	1.72	1.64	1.58	1.50	1.41	1.32	1.13
	0.01	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.06	1.90	1.79	1.72	1.61	1.50	1.38	1.16

TABLA 5 PROBABILIDADES BINOMIALES

Las entradas que aparecen en la tabla dan la probabilidad de x éxitos en n ensayos en un experimento binomial, donde p es la probabilidad de un éxito en un ensayo. Por ejemplo, para seis ensayos y $p = 0.05$, la probabilidad de dos éxitos es 0.0305.

		p								
n	***x***	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
2	0	0.9801	0.9604	0.9409	0.9216	0.9025	0.8836	0.8649	0.8464	0.8281
	1	0.0198	0.0392	0.0582	0.0768	0.0950	0.1128	0.1302	0.1472	0.1638
	2	0.0001	0.0004	0.0009	0.0016	0.0025	0.0036	0.0049	0.0064	0.0081
3	0	0.9703	0.9412	0.9127	0.8847	0.8574	0.8306	0.8044	0.7787	0.7536
	1	0.0294	0.0576	0.0847	0.1106	0.1354	0.1590	0.1816	0.2031	0.2236
	2	0.0003	0.0012	0.0026	0.0046	0.0071	0.0102	0.0137	0.0177	0.0221
	3	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003	0.0005	0.0007
4	0	0.9606	0.9224	0.8853	0.8493	0.8145	0.7807	0.7481	0.7164	0.6857
	1	0.0388	0.0753	0.1095	0.1416	0.1715	0.1993	0.2252	0.2492	0.2713
	2	0.0006	0.0023	0.0051	0.0088	0.0135	0.0191	0.0254	0.0325	0.0402
	3	0.0000	0.0000	0.0001	0.0002	0.0005	0.0008	0.0013	0.0019	0.0027
	4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
5	0	0.9510	0.9039	0.8587	0.8154	0.7738	0.7339	0.6957	0.6591	0.6240
	1	0.0480	0.0922	0.1328	0.1699	0.2036	0.2342	0.2618	0.2866	0.3086
	2	0.0010	0.0038	0.0082	0.0142	0.0214	0.0299	0.0394	0.0498	0.0610
	3	0.0000	0.0001	0.0003	0.0006	0.0011	0.0019	0.0030	0.0043	0.0060
	4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002	0.0003
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0	0.9415	0.8858	0.8330	0.7828	0.7351	0.6899	0.6470	0.6064	0.5679
	1	0.0571	0.1085	0.1546	0.1957	0.2321	0.2642	0.2922	0.3164	0.3370
	2	0.0014	0.0055	0.0120	0.0204	0.0305	0.0422	0.0550	0.0688	0.0833
	3	0.0000	0.0002	0.0005	0.0011	0.0021	0.0036	0.0055	0.0080	0.0110
	4	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005	0.0008
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0	0.9321	0.8681	0.8080	0.7514	0.6983	0.6485	0.6017	0.5578	0.5168
	1	0.0659	0.1240	0.1749	0.2192	0.2573	0.2897	0.3170	0.3396	0.3578
	2	0.0020	0.0076	0.0162	0.0274	0.0406	0.0555	0.0716	0.0886	0.1061
	3	0.0000	0.0003	0.0008	0.0019	0.0036	0.0059	0.0090	0.0128	0.0175
	4	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0011	0.0017
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	0	0.9227	0.8508	0.7837	0.7214	0.6634	0.6096	0.5596	0.5132	0.4703
	1	0.0746	0.1389	0.1939	0.2405	0.2793	0.3113	0.3370	0.3570	0.3721
	2	0.0026	0.0099	0.0210	0.0351	0.0515	0.0695	0.0888	0.1087	0.1288
	3	0.0001	0.0004	0.0013	0.0029	0.0054	0.0089	0.0134	0.0189	0.0255
	4	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0013	0.0021	0.0031
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

TABLA 5 PROBABILIDADES BINOMIALES (*continuación*)

		p								
n	***x***	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
9	0	0.9135	0.8337	0.7602	0.6925	0.6302	0.5730	0.5204	0.4722	0.4279
	1	0.0830	0.1531	0.2116	0.2597	0.2985	0.3292	0.3525	0.3695	0.3809
	2	0.0034	0.0125	0.0262	0.0433	0.0629	0.0840	0.1061	0.1285	0.1507
	3	0.0001	0.0006	0.0019	0.0042	0.0077	0.0125	0.0186	0.0261	0.0348
	4	0.0000	0.0000	0.0001	0.0003	0.0006	0.0012	0.0021	0.0034	0.0052
	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	0	0.9044	0.8171	0.7374	0.6648	0.5987	0.5386	0.4840	0.4344	0.3894
	1	0.0914	0.1667	0.2281	0.2770	0.3151	0.3438	0.3643	0.3777	0.3851
	2	0.0042	0.0153	0.0317	0.0519	0.0746	0.0988	0.1234	0.1478	0.1714
	3	0.0001	0.0008	0.0026	0.0058	0.0105	0.0168	0.0248	0.0343	0.0452
	4	0.0000	0.0000	0.0001	0.0004	0.0010	0.0019	0.0033	0.0052	0.0078
	5	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0003	0.0005	0.0009
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
12	0	0.8864	0.7847	0.6938	0.6127	0.5404	0.4759	0.4186	0.3677	0.3225
	1	0.1074	0.1922	0.2575	0.3064	0.3413	0.3645	0.3781	0.3837	0.3827
	2	0.0060	0.0216	0.0438	0.0702	0.0988	0.1280	0.1565	0.1835	0.2082
	3	0.0002	0.0015	0.0045	0.0098	0.0173	0.0272	0.0393	0.0532	0.0686
	4	0.0000	0.0001	0.0003	0.0009	0.0021	0.0039	0.0067	0.0104	0.0153
	5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0008	0.0014	0.0024
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0003
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15	0	0.8601	0.7386	0.6333	0.5421	0.4633	0.3953	0.3367	0.2863	0.2430
	1	0.1303	0.2261	0.2938	0.3388	0.3658	0.3785	0.3801	0.3734	0.3605
	2	0.0092	0.0323	0.0636	0.0988	0.1348	0.1691	0.2003	0.2273	0.2496
	3	0.0004	0.0029	0.0085	0.0178	0.0307	0.0468	0.0653	0.0857	0.1070
	4	0.0000	0.0002	0.0008	0.0022	0.0049	0.0090	0.0148	0.0223	0.0317
	5	0.0000	0.0000	0.0001	0.0002	0.0006	0.0013	0.0024	0.0043	0.0069
	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0006	0.0011
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

TABLA 5 PROBABILIDADES BINOMIALES (*continuación*)

		p								
n	***x***	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
18	0	0.8345	0.6951	0.5780	0.4796	0.3972	0.3283	0.2708	0.2229	0.1831
	1	0.1517	0.2554	0.3217	0.3597	0.3763	0.3772	0.3669	0.3489	0.3260
	2	0.0130	0.0443	0.0846	0.1274	0.1683	0.2047	0.2348	0.2579	0.2741
	3	0.0007	0.0048	0.0140	0.0283	0.0473	0.0697	0.0942	0.1196	0.1446
	4	0.0000	0.0004	0.0016	0.0044	0.0093	0.0167	0.0266	0.0390	0.0536
	5	0.0000	0.0000	0.0001	0.0005	0.0014	0.0030	0.0056	0.0095	0.0148
	6	0.0000	0.0000	0.0000	0.0000	0.0002	0.0004	0.0009	0.0018	0.0032
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0005
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
20	0	0.8179	0.6676	0.5438	0.4420	0.3585	0.2901	0.2342	0.1887	0.1516
	1	0.1652	0.2725	0.3364	0.3683	0.3774	0.3703	0.3526	0.3282	0.3000
	2	0.0159	0.0528	0.0988	0.1458	0.1887	0.2246	0.2521	0.2711	0.2818
	3	0.0010	0.0065	0.0183	0.0364	0.0596	0.0860	0.1139	0.1414	0.1672
	4	0.0000	0.0006	0.0024	0.0065	0.0133	0.0233	0.0364	0.0523	0.0703
	5	0.0000	0.0000	0.0002	0.0009	0.0022	0.0048	0.0088	0.0145	0.0222
	6	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0017	0.0032	0.0055
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0005	0.0011
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

TABLA 5 PROBABILIDADES BINOMIALES (*continuación*)

		p								
n	***x***	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
2	0	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000
	2	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500
3	0	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750
	2	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750
	3	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250
4	0	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500
	2	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750
	3	0.0036	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500
	4	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625
5	0	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0312
	1	0.3280	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1562
	2	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125
	3	0.0081	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125
	4	0.0004	0.0022	0.0064	0.0146	0.0284	0.0488	0.0768	0.1128	0.1562
	5	0.0000	0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0312
6	0	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.3543	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938
	2	0.0984	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344
	3	0.0146	0.0415	0.0819	0.1318	0.1852	0.2355	0.2765	0.3032	0.3125
	4	0.0012	0.0055	0.0154	0.0330	0.0595	0.0951	0.1382	0.1861	0.2344
	5	0.0001	0.0004	0.0015	0.0044	0.0102	0.0205	0.0369	0.0609	0.0938
	6	0.0000	0.0000	0.0001	0.0002	0.0007	0.0018	0.0041	0.0083	0.0156
7	0	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.3720	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547
	2	0.1240	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641
	3	0.0230	0.0617	0.1147	0.1730	0.2269	0.2679	0.2903	0.2918	0.2734
	4	0.0026	0.0109	0.0287	0.0577	0.0972	0.1442	0.1935	0.2388	0.2734
	5	0.0002	0.0012	0.0043	0.0115	0.0250	0.0466	0.0774	0.1172	0.1641
	6	0.0000	0.0001	0.0004	0.0013	0.0036	0.0084	0.0172	0.0320	0.0547
	7	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0016	0.0037	0.0078
8	0	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.3826	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0312
	2	0.1488	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094
	3	0.0331	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2188
	4	0.0046	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734
	5	0.0004	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2188
	6	0.0000	0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094
	7	0.0000	0.0000	0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0313
	8	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0017	0.0039

TABLA 5 PROBABILIDADES BINOMIALES (*continuación*)

		p								
n	***x***	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
9	0	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.3874	0.3679	0.3020	0.2253	0.1556	0.1004	0.0605	0.0339	0.0176
	2	0.1722	0.2597	0.3020	0.3003	0.2668	0.2162	0.1612	0.1110	0.0703
	3	0.0446	0.1069	0.1762	0.2336	0.2668	0.2716	0.2508	0.2119	0.1641
	4	0.0074	0.0283	0.0661	0.1168	0.1715	0.2194	0.2508	0.2600	0.2461
	5	0.0008	0.0050	0.0165	0.0389	0.0735	0.1181	0.1672	0.2128	0.2461
	6	0.0001	0.0006	0.0028	0.0087	0.0210	0.0424	0.0743	0.1160	0.1641
	7	0.0000	0.0000	0.0003	0.0012	0.0039	0.0098	0.0212	0.0407	0.0703
	8	0.0000	0.0000	0.0000	0.0001	0.0004	0.0013	0.0035	0.0083	0.0176
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0020
10	0	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1	0.3874	0.3474	0.2684	0.1877	0.1211	0.0725	0.0403	0.0207	0.0098
	2	0.1937	0.2759	0.3020	0.2816	0.2335	0.1757	0.1209	0.0763	0.0439
	3	0.0574	0.1298	0.2013	0.2503	0.2668	0.2522	0.2150	0.1665	0.1172
	4	0.0112	0.0401	0.0881	0.1460	0.2001	0.2377	0.2508	0.2384	0.2051
	5	0.0015	0.0085	0.0264	0.0584	0.1029	0.1536	0.2007	0.2340	0.2461
	6	0.0001	0.0012	0.0055	0.0162	0.0368	0.0689	0.1115	0.1596	0.2051
	7	0.0000	0.0001	0.0008	0.0031	0.0090	0.0212	0.0425	0.0746	0.1172
	8	0.0000	0.0000	0.0001	0.0004	0.0014	0.0043	0.0106	0.0229	0.0439
	9	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0016	0.0042	0.0098
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010
12	0	0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002
	1	0.3766	0.3012	0.2062	0.1267	0.0712	0.0368	0.0174	0.0075	0.0029
	2	0.2301	0.2924	0.2835	0.2323	0.1678	0.1088	0.0639	0.0339	0.0161
	3	0.0853	0.1720	0.2362	0.2581	0.2397	0.1954	0.1419	0.0923	0.0537
	4	0.0213	0.0683	0.1329	0.1936	0.2311	0.2367	0.2128	0.1700	0.1208
	5	0.0038	0.0193	0.0532	0.1032	0.1585	0.2039	0.2270	0.2225	0.1934
	6	0.0005	0.0040	0.0155	0.0401	0.0792	0.1281	0.1766	0.2124	0.2256
	7	0.0000	0.0006	0.0033	0.0115	0.0291	0.0591	0.1009	0.1489	0.1934
	8	0.0000	0.0001	0.0005	0.0024	0.0078	0.0199	0.0420	0.0762	0.1208
	9	0.0000	0.0000	0.0001	0.0004	0.0015	0.0048	0.0125	0.0277	0.0537
	10	0.0000	0.0000	0.0000	0.0000	0.0002	0.0008	0.0025	0.0068	0.0161
	11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010	0.0029
	12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002
15	0	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000
	1	0.3432	0.2312	0.1319	0.0668	0.0305	0.0126	0.0047	0.0016	0.0005
	2	0.2669	0.2856	0.2309	0.1559	0.0916	0.0476	0.0219	0.0090	0.0032
	3	0.1285	0.2184	0.2501	0.2252	0.1700	0.1110	0.0634	0.0318	0.0139
	4	0.0428	0.1156	0.1876	0.2252	0.2186	0.1792	0.1268	0.0780	0.0417
	5	0.0105	0.0449	0.1032	0.1651	0.2061	0.2123	0.1859	0.1404	0.0916
	6	0.0019	0.0132	0.0430	0.0917	0.1472	0.1906	0.2066	0.1914	0.1527
	7	0.0003	0.0030	0.0138	0.0393	0.0811	0.1319	0.1771	0.2013	0.1964
	8	0.0000	0.0005	0.0035	0.0131	0.0348	0.0710	0.1181	0.1647	0.1964
	9	0.0000	0.0001	0.0007	0.0034	0.0016	0.0298	0.0612	0.1048	0.1527
	10	0.0000	0.0000	0.0001	0.0007	0.0030	0.0096	0.0245	0.0515	0.0916
	11	0.0000	0.0000	0.0000	0.0001	0.0006	0.0024	0.0074	0.0191	0.0417
	12	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0016	0.0052	0.0139
	13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010	0.0032
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

TABLA 5 PROBABILIDADES BINOMIALES (*continuación*)

		p								
n	***x***	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
18	0	0.1501	0.0536	0.0180	0.0056	0.0016	0.0004	0.0001	0.0000	0.0000
	1	0.3002	0.1704	0.0811	0.0338	0.0126	0.0042	0.0012	0.0003	0.0001
	2	0.2835	0.2556	0.1723	0.0958	0.0458	0.0190	0.0069	0.0022	0.0006
	3	0.1680	0.2406	0.2297	0.1704	0.1046	0.0547	0.0246	0.0095	0.0031
	4	0.0700	0.1592	0.2153	0.2130	0.1681	0.1104	0.0614	0.0291	0.0117
	5	0.0218	0.0787	0.1507	0.1988	0.2017	0.1664	0.1146	0.0666	0.0327
	6	0.0052	0.0301	0.0816	0.1436	0.1873	0.1941	0.1655	0.1181	0.0708
	7	0.0010	0.0091	0.0350	0.0820	0.1376	0.1792	0.1892	0.1657	0.1214
	8	0.0002	0.0022	0.0120	0.0376	0.0811	0.1327	0.1734	0.1864	0.1669
	9	0.0000	0.0004	0.0033	0.0139	0.0386	0.0794	0.1284	0.1694	0.1855
	10	0.0000	0.0001	0.0008	0.0042	0.0149	0.0385	0.0771	0.1248	0.1669
	11	0.0000	0.0000	0.0001	0.0010	0.0046	0.0151	0.0374	0.0742	0.1214
	12	0.0000	0.0000	0.0000	0.0002	0.0012	0.0047	0.0145	0.0354	0.0708
	13	0.0000	0.0000	0.0000	0.0000	0.0002	0.0012	0.0045	0.0134	0.0327
	14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0011	0.0039	0.0117
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0009	0.0031
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
20	0	0.1216	0.0388	0.0115	0.0032	0.0008	0.0002	0.0000	0.0000	0.0000
	1	0.2702	0.1368	0.0576	0.0211	0.0068	0.0020	0.0005	0.0001	0.0000
	2	0.2852	0.2293	0.1369	0.0669	0.0278	0.0100	0.0031	0.0008	0.0002
	3	0.1901	0.2428	0.2054	0.1339	0.0716	0.0323	0.0123	0.0040	0.0011
	4	0.0898	0.1821	0.2182	0.1897	0.1304	0.0738	0.0350	0.0139	0.0046
	5	0.0319	0.1028	0.1746	0.2023	0.1789	0.1272	0.0746	0.0365	0.0148
	6	0.0089	0.0454	0.1091	0.1686	0.1916	0.1712	0.1244	0.0746	0.0370
	7	0.0020	0.0160	0.0545	0.1124	0.1643	0.1844	0.1659	0.1221	0.0739
	8	0.0004	0.0046	0.0222	0.0609	0.1144	0.1614	0.1797	0.1623	0.1201
	9	0.0001	0.0011	0.0074	0.0271	0.0654	0.1158	0.1597	0.1771	0.1602
	10	0.0000	0.0002	0.0020	0.0099	0.0308	0.0686	0.1171	0.1593	0.1762
	11	0.0000	0.0000	0.0005	0.0030	0.0120	0.0336	0.0710	0.1185	0.1602
	12	0.0000	0.0000	0.0001	0.0008	0.0039	0.0136	0.0355	0.0727	0.1201
	13	0.0000	0.0000	0.0000	0.0002	0.0010	0.0045	0.0146	0.0366	0.0739
	14	0.0000	0.0000	0.0000	0.0000	0.0002	0.0012	0.0049	0.0150	0.0370
	15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0013	0.0049	0.0148
	16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0013	0.0046
	17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0011
	18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

TABLA 5 PROBABILIDADES BINOMIALES (*continuación*)

		p								
n	***x***	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
2	0	0.2025	0.1600	0.1225	0.0900	0.0625	0.0400	0.0225	0.0100	0.0025
	1	0.4950	0.4800	0.4550	0.4200	0.3750	0.3200	0.2550	0.1800	0.0950
	2	0.3025	0.3600	0.4225	0.4900	0.5625	0.6400	0.7225	0.8100	0.9025
3	0	0.0911	0.0640	0.0429	0.0270	0.0156	0.0080	0.0034	0.0010	0.0001
	1	0.3341	0.2880	0.2389	0.1890	0.1406	0.0960	0.0574	0.0270	0.0071
	2	0.4084	0.4320	0.4436	0.4410	0.4219	0.3840	0.3251	0.2430	0.1354
	3	0.1664	0.2160	0.2746	0.3430	0.4219	0.5120	0.6141	0.7290	0.8574
4	0	0.0410	0.0256	0.0150	0.0081	0.0039	0.0016	0.0005	0.0001	0.0000
	1	0.2005	0.1536	0.1115	0.0756	0.0469	0.0256	0.0115	0.0036	0.0005
	2	0.3675	0.3456	0.3105	0.2646	0.2109	0.1536	0.0975	0.0486	0.0135
	3	0.2995	0.3456	0.3845	0.4116	0.4219	0.4096	0.3685	0.2916	0.1715
	4	0.0915	0.1296	0.1785	0.2401	0.3164	0.4096	0.5220	0.6561	0.8145
5	0	0.0185	0.0102	0.0053	0.0024	0.0010	0.0003	0.0001	0.0000	0.0000
	1	0.1128	0.0768	0.0488	0.0284	0.0146	0.0064	0.0022	0.0005	0.0000
	2	0.2757	0.2304	0.1811	0.1323	0.0879	0.0512	0.0244	0.0081	0.0011
	3	0.3369	0.3456	0.3364	0.3087	0.2637	0.2048	0.1382	0.0729	0.0214
	4	0.2059	0.2592	0.3124	0.3601	0.3955	0.4096	0.3915	0.3281	0.2036
	5	0.0503	0.0778	0.1160	0.1681	0.2373	0.3277	0.4437	0.5905	0.7738
6	0	0.0083	0.0041	0.0018	0.0007	0.0002	0.0001	0.0000	0.0000	0.0000
	1	0.0609	0.0369	0.0205	0.0102	0.0044	0.0015	0.0004	0.0001	0.0000
	2	0.1861	0.1382	0.0951	0.0595	0.0330	0.0154	0.0055	0.0012	0.0001
	3	0.3032	0.2765	0.2355	0.1852	0.1318	0.0819	0.0415	0.0146	0.0021
	4	0.2780	0.3110	0.3280	0.3241	0.2966	0.2458	0.1762	0.0984	0.0305
	5	0.1359	0.1866	0.2437	0.3025	0.3560	0.3932	0.3993	0.3543	0.2321
	6	0.0277	0.0467	0.0754	0.1176	0.1780	0.2621	0.3771	0.5314	0.7351
7	0	0.0037	0.0016	0.0006	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
	1	0.0320	0.0172	0.0084	0.0036	0.0013	0.0004	0.0001	0.0000	0.0000
	2	0.1172	0.0774	0.0466	0.0250	0.0115	0.0043	0.0012	0.0002	0.0000
	3	0.2388	0.1935	0.1442	0.0972	0.0577	0.0287	0.0109	0.0026	0.0002
	4	0.2918	0.2903	0.2679	0.2269	0.1730	0.1147	0.0617	0.0230	0.0036
	5	0.2140	0.2613	0.2985	0.3177	0.3115	0.2753	0.2097	0.1240	0.0406
	6	0.0872	0.1306	0.1848	0.2471	0.3115	0.3670	0.3960	0.3720	0.2573
	7	0.0152	0.0280	0.0490	0.0824	0.1335	0.2097	0.3206	0.4783	0.6983
8	0	0.0017	0.0007	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0164	0.0079	0.0033	0.0012	0.0004	0.0001	0.0000	0.0000	0.0000
	2	0.0703	0.0413	0.0217	0.0100	0.0038	0.0011	0.0002	0.0000	0.0000
	3	0.1719	0.1239	0.0808	0.0467	0.0231	0.0092	0.0026	0.0004	0.0000
	4	0.2627	0.2322	0.1875	0.1361	0.0865	0.0459	0.0185	0.0046	0.0004
	5	0.2568	0.2787	0.2786	0.2541	0.2076	0.1468	0.0839	0.0331	0.0054
	6	0.1569	0.2090	0.2587	0.2965	0.3115	0.2936	0.2376	0.1488	0.0515
	7	0.0548	0.0896	0.1373	0.1977	0.2670	0.3355	0.3847	0.3826	0.2793
	8	0.0084	0.0168	0.0319	0.0576	0.1001	0.1678	0.2725	0.4305	0.6634

TABLA 5 PROBABILIDADES BINOMIALES (*continuación*)

		p								
n	***x***	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
9	0	0.0008	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0083	0.0035	0.0013	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000
	2	0.0407	0.0212	0.0098	0.0039	0.0012	0.0003	0.0000	0.0000	0.0000
	3	0.1160	0.0743	0.0424	0.0210	0.0087	0.0028	0.0006	0.0001	0.0000
	4	0.2128	0.1672	0.1181	0.0735	0.0389	0.0165	0.0050	0.0008	0.0000
	5	0.2600	0.2508	0.2194	0.1715	0.1168	0.0661	0.0283	0.0074	0.0006
	6	0.2119	0.2508	0.2716	0.2668	0.2336	0.1762	0.1069	0.0446	0.0077
	7	0.1110	0.1612	0.2162	0.2668	0.3003	0.3020	0.2597	0.1722	0.0629
	8	0.0339	0.0605	0.1004	0.1556	0.2253	0.3020	0.3679	0.3874	0.2985
	9	0.0046	0.0101	0.0207	0.0404	0.0751	0.1342	0.2316	0.3874	0.6302
10	0	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0042	0.0016	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0229	0.0106	0.0043	0.0014	0.0004	0.0001	0.0000	0.0000	0.0000
	3	0.0746	0.0425	0.0212	0.0090	0.0031	0.0008	0.0001	0.0000	0.0000
	4	0.1596	0.1115	0.0689	0.0368	0.0162	0.0055	0.0012	0.0001	0.0000
	5	0.2340	0.2007	0.1536	0.1029	0.0584	0.0264	0.0085	0.0015	0.0001
	6	0.2384	0.2508	0.2377	0.2001	0.1460	0.0881	0.0401	0.0112	0.0010
	7	0.1665	0.2150	0.2522	0.2668	0.2503	0.2013	0.1298	0.0574	0.0105
	8	0.0763	0.1209	0.1757	0.2335	0.2816	0.3020	0.2759	0.1937	0.0746
	9	0.0207	0.0403	0.0725	0.1211	0.1877	0.2684	0.3474	0.3874	0.3151
	10	0.0025	0.0060	0.0135	0.0282	0.0563	0.1074	0.1969	0.3487	0.5987
12	0	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0010	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0068	0.0025	0.0008	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0277	0.0125	0.0048	0.0015	0.0004	0.0001	0.0000	0.0000	0.0000
	4	0.0762	0.0420	0.0199	0.0078	0.0024	0.0005	0.0001	0.0000	0.0000
	5	0.1489	0.1009	0.0591	0.0291	0.0115	0.0033	0.0006	0.0000	0.0000
	6	0.2124	0.1766	0.1281	0.0792	0.0401	0.0155	0.0040	0.0005	0.0000
	7	0.2225	0.2270	0.2039	0.1585	0.1032	0.0532	0.0193	0.0038	0.0002
	8	0.1700	0.2128	0.2367	0.2311	0.1936	0.1329	0.0683	0.0213	0.0021
	9	0.0923	0.1419	0.1954	0.2397	0.2581	0.2362	0.1720	0.0852	0.0173
	10	0.0339	0.0639	0.1088	0.1678	0.2323	0.2835	0.2924	0.2301	0.0988
	11	0.0075	0.0174	0.0368	0.0712	0.1267	0.2062	0.3012	0.3766	0.3413
	12	0.0008	0.0022	0.0057	0.0138	0.0317	0.0687	0.1422	0.2824	0.5404
15	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0010	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0052	0.0016	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0191	0.0074	0.0024	0.0006	0.0001	0.0000	0.0000	0.0000	0.0000
	5	0.0515	0.0245	0.0096	0.0030	0.0007	0.0001	0.0000	0.0000	0.0000
	6	0.1048	0.0612	0.0298	0.0116	0.0034	0.0007	0.0001	0.0000	0.0000
	7	0.1647	0.1181	0.0710	0.0348	0.0131	0.0035	0.0005	0.0000	0.0000
	8	0.2013	0.1771	0.1319	0.0811	0.0393	0.0138	0.0030	0.0003	0.0000
	9	0.1914	0.2066	0.1906	0.1472	0.0917	0.0430	0.0132	0.0019	0.0000
	10	0.1404	0.1859	0.2123	0.2061	0.1651	0.1032	0.0449	0.0105	0.0006
	11	0.0780	0.1268	0.1792	0.2186	0.2252	0.1876	0.1156	0.0428	0.0049

TABLA 5 PROBABILIDADES BINOMIALES (*continuación*)

		p								
n	***x***	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
	12	0.0318	0.0634	0.1110	0.1700	0.2252	0.2501	0.2184	0.1285	0.0307
	13	0.0090	0.0219	0.0476	0.0916	0.1559	0.2309	0.2856	0.2669	0.1348
	14	0.0016	0.0047	0.0126	0.0305	0.0668	0.1319	0.2312	0.3432	0.3658
	15	0.0001	0.0005	0.0016	0.0047	0.0134	0.0352	0.0874	0.2059	0.4633
18	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0009	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0039	0.0011	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.0134	0.0045	0.0012	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.0354	0.0145	0.0047	0.0012	0.0002	0.0000	0.0000	0.0000	0.0000
	7	0.0742	0.0374	0.0151	0.0046	0.0010	0.0001	0.0000	0.0000	0.0000
	8	0.1248	0.0771	0.0385	0.0149	0.0042	0.0008	0.0001	0.0000	0.0000
	9	0.1694	0.1284	0.0794	0.0386	0.0139	0.0033	0.0004	0.0000	0.0000
	10	0.1864	0.1734	0.1327	0.0811	0.0376	0.0120	0.0022	0.0002	0.0000
	11	0.1657	0.1892	0.1792	0.1376	0.0820	0.0350	0.0091	0.0010	0.0000
	12	0.1181	0.1655	0.1941	0.1873	0.1436	0.0816	0.0301	0.0052	0.0002
	13	0.0666	0.1146	0.1664	0.2017	0.1988	0.1507	0.0787	0.0218	0.0014
	14	0.0291	0.0614	0.1104	0.1681	0.2130	0.2153	0.1592	0.0700	0.0093
	15	0.0095	0.0246	0.0547	0.1046	0.1704	0.2297	0.2406	0.1680	0.0473
	16	0.0022	0.0069	0.0190	0.0458	0.0958	0.1723	0.2556	0.2835	0.1683
	17	0.0003	0.0012	0.0042	0.0126	0.0338	0.0811	0.1704	0.3002	0.3763
	18	0.0000	0.0001	0.0004	0.0016	0.0056	0.0180	0.0536	0.1501	0.3972
20	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	0.0013	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	5	0.0049	0.0013	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	6	0.0150	0.0049	0.0012	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0366	0.0146	0.0045	0.0010	0.0002	0.0000	0.0000	0.0000	0.0000
	8	0.0727	0.0355	0.0136	0.0039	0.0008	0.0001	0.0000	0.0000	0.0000
	9	0.1185	0.0710	0.0336	0.0120	0.0030	0.0005	0.0000	0.0000	0.0000
	10	0.1593	0.1171	0.0686	0.0308	0.0099	0.0020	0.0002	0.0000	0.0000
	11	0.1771	0.1597	0.1158	0.0654	0.0271	0.0074	0.0011	0.0001	0.0000
	12	0.1623	0.1797	0.1614	0.1144	0.0609	0.0222	0.0046	0.0004	0.0000
	13	0.1221	0.1659	0.1844	0.1643	0.1124	0.0545	0.0160	0.0020	0.0000
	14	0.0746	0.1244	0.1712	0.1916	0.1686	0.1091	0.0454	0.0089	0.0003
	15	0.0365	0.0746	0.1272	0.1789	0.2023	0.1746	0.1028	0.0319	0.0022
	16	0.0139	0.0350	0.0738	0.1304	0.1897	0.2182	0.1821	0.0898	0.0133
	17	0.0040	0.0123	0.0323	0.0716	0.1339	0.2054	0.2428	0.1901	0.0596
	18	0.0008	0.0031	0.0100	0.0278	0.0669	0.1369	0.2293	0.2852	0.1887
	19	0.0001	0.0005	0.0020	0.0068	0.0211	0.0576	0.1368	0.2702	0.3774
	20	0.0000	0.0000	0.0002	0.0008	0.0032	0.0115	0.0388	0.1216	0.3585

TABLA 6 VALORES DE $e^{-\mu}$

μ	$e^{-\mu}$	μ	$e^{-\mu}$	μ	$e^{-\mu}$
0.00	1.0000	2.00	0.1353	4.00	0.0183
0.05	0.9512	2.05	0.1287	4.05	0.0174
0.10	0.9048	2.10	0.1225	4.10	0.0166
0.15	0.8607	2.15	0.1165	4.15	0.0158
0.20	0.8187	2.20	0.1108	4.20	0.0150
0.25	0.7788	2.25	0.1054	4.25	0.0143
0.30	0.7408	2.30	0.1003	4.30	0.0136
0.35	0.7047	2.35	0.0954	4.35	0.0129
0.40	0.6703	2.40	0.0907	4.40	0.0123
0.45	0.6376	2.45	0.0863	4.45	0.0117
0.50	0.6065	2.50	0.0821	4.50	0.0111
0.55	0.5769	2.55	0.0781	4.55	0.0106
0.60	0.5488	2.60	0.0743	4.60	0.0101
0.65	0.5220	2.65	0.0707	4.65	0.0096
0.70	0.4966	2.70	0.0672	4.70	0.0091
0.75	0.4724	2.75	0.0639	4.75	0.0087
0.80	0.4493	2.80	0.0608	4.80	0.0082
0.85	0.4274	2.85	0.0578	4.85	0.0078
0.90	0.4066	2.90	0.0550	4.90	0.0074
0.95	0.3867	2.95	0.0523	4.95	0.0071
1.00	0.3679	3.00	0.0498	5.00	0.0067
1.05	0.3499	3.05	0.0474	6.00	0.0025
1.10	0.3329	3.10	0.0450	7.00	0.0009
1.15	0.3166	3.15	0.0429	8.00	0.000335
1.20	0.3012	3.20	0.0408	9.00	0.000123
				10.00	0.000045
1.25	0.2865	3.25	0.0388		
1.30	0.2725	3.30	0.0369		
1.35	0.2592	3.35	0.0351		
1.40	0.2466	3.40	0.0334		
1.45	0.2346	3.45	0.0317		
1.50	0.2231	3.50	0.0302		
1.55	0.2122	3.55	0.0287		
1.60	0.2019	3.60	0.0273		
1.65	0.1920	3.65	0.0260		
1.70	0.1827	3.70	0.0247		
1.75	0.1738	3.75	0.0235		
1.80	0.1653	3.80	0.0224		
1.85	0.1572	3.85	0.0213		
1.90	0.1496	3.90	0.0202		
1.95	0.1423	3.95	0.0193		

TABLA 7 PROBABILIDADES POISSON

Las entradas que aparecen en la tabla dan la probabilidad de x ocurrencias en un proceso de Poisson cuya media es μ. Por ejemplo, si $\mu = 2.5$, la probabilidad de cuatro ocurrencias es 0.1336.

	μ									
x	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659	0.3679
2	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647	0.1839
3	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494	0.0613
4	0.0000	0.0001	0.0002	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111	0.0153
5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0012	0.0020	0.0031
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

	μ									
x	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**1.8**	**1.9**	**2.0**
0	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496	0.1353
1	0.3662	0.3614	0.3543	0.3452	0.3347	0.3230	0.3106	0.2975	0.2842	0.2707
2	0.2014	0.2169	0.2303	0.2417	0.2510	0.2584	0.2640	0.2678	0.2700	0.2707
3	0.0738	0.0867	0.0998	0.1128	0.1255	0.1378	0.1496	0.1607	0.1710	0.1804
4	0.0203	0.0260	0.0324	0.0395	0.0471	0.0551	0.0636	0.0723	0.0812	0.0902
5	0.0045	0.0062	0.0084	0.0111	0.0141	0.0176	0.0216	0.0260	0.0309	0.0361
6	0.0008	0.0012	0.0018	0.0026	0.0035	0.0047	0.0061	0.0078	0.0098	0.0120
7	0.0001	0.0002	0.0003	0.0005	0.0008	0.0011	0.0015	0.0020	0.0027	0.0034
8	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0003	0.0005	0.0006	0.0009
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002

	μ									
x	**2.1**	**2.2**	**2.3**	**2.4**	**2.5**	**2.6**	**2.7**	**2.8**	**2.9**	**3.0**
0	0.1225	0.1108	0.1003	0.0907	0.0821	0.0743	0.0672	0.0608	0.0550	0.0498
1	0.2572	0.2438	0.2306	0.2177	0.2052	0.1931	0.1815	0.1703	0.1596	0.1494
2	0.2700	0.2681	0.2652	0.2613	0.2565	0.2510	0.2450	0.2384	0.2314	0.2240
3	0.1890	0.1966	0.2033	0.2090	0.2138	0.2176	0.2205	0.2225	0.2237	0.2240
4	0.0992	0.1082	0.1169	0.1254	0.1336	0.1414	0.1488	0.1557	0.1622	0.1680
5	0.0417	0.0476	0.0538	0.0602	0.0668	0.0735	0.0804	0.0872	0.0940	0.1008
6	0.0146	0.0174	0.0206	0.0241	0.0278	0.0319	0.0362	0.0407	0.0455	0.0504
7	0.0044	0.0055	0.0068	0.0083	0.0099	0.0118	0.0139	0.0163	0.0188	0.0216
8	0.0011	0.0015	0.0019	0.0025	0.0031	0.0038	0.0047	0.0057	0.0068	0.0081
9	0.0003	0.0004	0.0005	0.0007	0.0009	0.0011	0.0014	0.0018	0.0022	0.0027
10	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0004	0.0005	0.0006	0.0008
11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0002
12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

TABLA 7 PROBABILIDADES POISSON (*continuación*)

	μ									
x	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**	**3.7**	**3.8**	**3.9**	**4.0**
0	0.0450	0.0408	0.0369	0.0344	0.0302	0.0273	0.0247	0.0224	0.0202	0.0183
1	0.1397	0.1304	0.1217	0.1135	0.1057	0.0984	0.0915	0.0850	0.0789	0.0733
2	0.2165	0.2087	0.2008	0.1929	0.1850	0.1771	0.1692	0.1615	0.1539	0.1465
3	0.2237	0.2226	0.2209	0.2186	0.2158	0.2125	0.2087	0.2046	0.2001	0.1954
4	0.1734	0.1781	0.1823	0.1858	0.1888	0.1912	0.1931	0.1944	0.1951	0.1954
5	0.1075	0.1140	0.1203	0.1264	0.1322	0.1377	0.1429	0.1477	0.1522	0.1563
6	0.0555	0.0608	0.0662	0.0716	0.0771	0.0826	0.0881	0.0936	0.0989	0.1042
7	0.0246	0.0278	0.0312	0.0348	0.0385	0.0425	0.0466	0.0508	0.0551	0.0595
8	0.0095	0.0111	0.0129	0.0148	0.0169	0.0191	0.0215	0.0241	0.0269	0.0298
9	0.0033	0.0040	0.0047	0.0056	0.0066	0.0076	0.0089	0.0102	0.0116	0.0132
10	0.0010	0.0013	0.0016	0.0019	0.0023	0.0028	0.0033	0.0039	0.0045	0.0053
11	0.0003	0.0004	0.0005	0.0006	0.0007	0.0009	0.0011	0.0013	0.0016	0.0019
12	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005	0.0006
13	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

	μ									
x	**4.1**	**4.2**	**4.3**	**4.4**	**4.5**	**4.6**	**4.7**	**4.8**	**4.9**	**5.0**
0	0.0166	0.0150	0.0136	0.0123	0.0111	0.0101	0.0091	0.0082	0.0074	0.0067
1	0.0679	0.0630	0.0583	0.0540	0.0500	0.0462	0.0427	0.0395	0.0365	0.0337
2	0.1393	0.1323	0.1254	0.1188	0.1125	0.1063	0.1005	0.0948	0.0894	0.0842
3	0.1904	0.1852	0.1798	0.1743	0.1687	0.1631	0.1574	0.1517	0.1460	0.1404
4	0.1951	0.1944	0.1933	0.1917	0.1898	0.1875	0.1849	0.1820	0.1789	0.1755
5	0.1600	0.1633	0.1662	0.1687	0.1708	0.1725	0.1738	0.1747	0.1753	0.1755
6	0.1093	0.1143	0.1191	0.1237	0.1281	0.1323	0.1362	0.1398	0.1432	0.1462
7	0.0640	0.0686	0.0732	0.0778	0.0824	0.0869	0.0914	0.0959	0.1002	0.1044
8	0.0328	0.0360	0.0393	0.0428	0.0463	0.0500	0.0537	0.0575	0.0614	0.0653
9	0.0150	0.0168	0.0188	0.0209	0.0232	0.0255	0.0280	0.0307	0.0334	0.0363
10	0.0061	0.0071	0.0081	0.0092	0.0104	0.0118	0.0132	0.0147	0.0164	0.0181
11	0.0023	0.0027	0.0032	0.0037	0.0043	0.0049	0.0056	0.0064	0.0073	0.0082
12	0.0008	0.0009	0.0011	0.0014	0.0016	0.0019	0.0022	0.0026	0.0030	0.0034
13	0.0002	0.0003	0.0004	0.0005	0.0006	0.0007	0.0008	0.0009	0.0011	0.0013
14	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005
15	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002

	μ									
x	**5.1**	**5.2**	**5.3**	**5.4**	**5.5**	**5.6**	**5.7**	**5.8**	**5.9**	**6.0**
0	0.0061	0.0055	0.0050	0.0045	0.0041	0.0037	0.0033	0.0030	0.0027	0.0025
1	0.0311	0.0287	0.0265	0.0244	0.0225	0.0207	0.0191	0.0176	0.0162	0.0149
2	0.0793	0.0746	0.0701	0.0659	0.0618	0.0580	0.0544	0.0509	0.0477	0.0446
3	0.1348	0.1293	0.1239	0.1185	0.1133	0.1082	0.1033	0.0985	0.0938	0.0892
4	0.1719	0.1681	0.1641	0.1600	0.1558	0.1515	0.1472	0.1428	0.1383	0.1339

TABLA 7 PROBABILIDADES POISSON (*continuación*)

	μ									
x	**5.1**	**5.2**	**5.3**	**5.4**	**5.5**	**5.6**	**5.7**	**5.8**	**5.9**	**6.0**
5	0.1753	0.1748	0.1740	0.1728	0.1714	0.1697	0.1678	0.1656	0.1632	0.1606
6	0.1490	0.1515	0.1537	0.1555	0.1571	0.1587	0.1594	0.1601	0.1605	0.1606
7	0.1086	0.1125	0.1163	0.1200	0.1234	0.1267	0.1298	0.1326	0.1353	0.1377
8	0.0692	0.0731	0.0771	0.0810	0.0849	0.0887	0.0925	0.0962	0.0998	0.1033
9	0.0392	0.0423	0.0454	0.0486	0.0519	0.0552	0.0586	0.0620	0.0654	0.0688
10	0.0200	0.0220	0.0241	0.0262	0.0285	0.0309	0.0334	0.0359	0.0386	0.0413
11	0.0093	0.0104	0.0116	0.0129	0.0143	0.0157	0.0173	0.0190	0.0207	0.0225
12	0.0039	0.0045	0.0051	0.0058	0.0065	0.0073	0.0082	0.0092	0.0102	0.0113
13	0.0015	0.0018	0.0021	0.0024	0.0028	0.0032	0.0036	0.0041	0.0046	0.0052
14	0.0006	0.0007	0.0008	0.0009	0.0011	0.0013	0.0015	0.0017	0.0019	0.0022
15	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005	0.0006	0.0007	0.0008	0.0009
16	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003
17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001

	μ									
x	**6.1**	**6.2**	**6.3**	**6.4**	**6.5**	**6.6**	**6.7**	**6.8**	**6.9**	**7.0**
0	0.0022	0.0020	0.0018	0.0017	0.0015	0.0014	0.0012	0.0011	0.0010	0.0009
1	0.0137	0.0126	0.0116	0.0106	0.0098	0.0090	0.0082	0.0076	0.0070	0.0064
2	0.0417	0.0390	0.0364	0.0340	0.0318	0.0296	0.0276	0.0258	0.0240	0.0223
3	0.0848	0.0806	0.0765	0.0726	0.0688	0.0652	0.0617	0.0584	0.0552	0.0521
4	0.1294	0.1249	0.1205	0.1162	0.1118	0.1076	0.1034	0.0992	0.0952	0.0912
5	0.1579	0.1549	0.1519	0.1487	0.1454	0.1420	0.1385	0.1349	0.1314	0.1277
6	0.1605	0.1601	0.1595	0.1586	0.1575	0.1562	0.1546	0.1529	0.1511	0.1490
7	0.1399	0.1418	0.1435	0.1450	0.1462	0.1472	0.1480	0.1486	0.1489	0.1490
8	0.1066	0.1099	0.1130	0.1160	0.1188	0.1215	0.1240	0.1263	0.1284	0.1304
9	0.0723	0.0757	0.0791	0.0825	0.0858	0.0891	0.0923	0.0954	0.0985	0.1014
10	0.0441	0.0469	0.0498	0.0528	0.0558	0.0588	0.0618	0.0649	0.0679	0.0710
11	0.0245	0.0265	0.0285	0.0307	0.0330	0.0353	0.0377	0.0401	0.0426	0.0452
12	0.0124	0.0137	0.0150	0.0164	0.0179	0.0194	0.0210	0.0227	0.0245	0.0264
13	0.0058	0.0065	0.0073	0.0081	0.0089	0.0098	0.0108	0.0119	0.0130	0.0142
14	0.0025	0.0029	0.0033	0.0037	0.0041	0.0046	0.0052	0.0058	0.0064	0.0071
15	0.0010	0.0012	0.0014	0.0016	0.0018	0.0020	0.0023	0.0026	0.0029	0.0033
16	0.0004	0.0005	0.0005	0.0006	0.0007	0.0008	0.0010	0.0011	0.0013	0.0014
17	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006
18	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002
19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001

	μ									
x	**7.1**	**7.2**	**7.3**	**7.4**	**7.5**	**7.6**	**7.7**	**7.8**	**7.9**	**8.0**
0	0.0008	0.0007	0.0007	0.0006	0.0006	0.0005	0.0005	0.0004	0.0004	0.0003
1	0.0059	0.0054	0.0049	0.0045	0.0041	0.0038	0.0035	0.0032	0.0029	0.0027
2	0.0208	0.0194	0.0180	0.0167	0.0156	0.0145	0.0134	0.0125	0.0116	0.0107
3	0.0492	0.0464	0.0438	0.0413	0.0389	0.0366	0.0345	0.0324	0.0305	0.0286
4	0.0874	0.0836	0.0799	0.0764	0.0729	0.0696	0.0663	0.0632	0.0602	0.0573

TABLA 7 PROBABILIDADES POISSON (*continuación*)

	μ									
x	**7.1**	**7.2**	**7.3**	**7.4**	**7.5**	**7.6**	**7.7**	**7.8**	**7.9**	**8.0**
5	0.1241	0.1204	0.1167	0.1130	0.1094	0.1057	0.1021	0.0986	0.0951	0.0916
6	0.1468	0.1445	0.1420	0.1394	0.1367	0.1339	0.1311	0.1282	0.1252	0.1221
7	0.1489	0.1486	0.1481	0.1474	0.1465	0.1454	0.1442	0.1428	0.1413	0.1396
8	0.1321	0.1337	0.1351	0.1363	0.1373	0.1382	0.1388	0.1392	0.1395	0.1396
9	0.1042	0.1070	0.1096	0.1121	0.1144	0.1167	0.1187	0.1207	0.1224	0.1241
10	0.0740	0.0770	0.0800	0.0829	0.0858	0.0887	0.0914	0.0941	0.0967	0.0993
11	0.0478	0.0504	0.0531	0.0558	0.0585	0.0613	0.0640	0.0667	0.0695	0.0722
12	0.0283	0.0303	0.0323	0.0344	0.0366	0.0388	0.0411	0.0434	0.0457	0.0481
13	0.0154	0.0168	0.0181	0.0196	0.0211	0.0227	0.0243	0.0260	0.0278	0.0296
14	0.0078	0.0086	0.0095	0.0104	0.0113	0.0123	0.0134	0.0145	0.0157	0.0169
15	0.0037	0.0041	0.0046	0.0051	0.0057	0.0062	0.0069	0.0075	0.0083	0.0090
16	0.0016	0.0019	0.0021	0.0024	0.0026	0.0030	0.0033	0.0037	0.0041	0.0045
17	0.0007	0.0008	0.0009	0.0010	0.0012	0.0013	0.0015	0.0017	0.0019	0.0021
18	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0007	0.0008	0.0009
19	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003	0.0003	0.0004
20	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002
21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001

	μ									
x	**8.1**	**8.2**	**8.3**	**8.4**	**8.5**	**8.6**	**8.7**	**8.8**	**8.9**	**9.0**
0	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001
1	0.0025	0.0023	0.0021	0.0019	0.0017	0.0016	0.0014	0.0013	0.0012	0.0011
2	0.0100	0.0092	0.0086	0.0079	0.0074	0.0068	0.0063	0.0058	0.0054	0.0050
3	0.0269	0.0252	0.0237	0.0222	0.0208	0.0195	0.0183	0.0171	0.0160	0.0150
4	0.0544	0.0517	0.0491	0.0466	0.0443	0.0420	0.0398	0.0377	0.0357	0.0337
5	0.0882	0.0849	0.0816	0.0784	0.0752	0.0722	0.0692	0.0663	0.0635	0.0607
6	0.1191	0.1160	0.1128	0.1097	0.1066	0.1034	0.1003	0.0972	0.0941	0.0911
7	0.1378	0.1358	0.1338	0.1317	0.1294	0.1271	0.1247	0.1222	0.1197	0.1171
8	0.1395	0.1392	0.1388	0.1382	0.1375	0.1366	0.1356	0.1344	0.1332	0.1318
9	0.1256	0.1269	0.1280	0.1290	0.1299	0.1306	0.1311	0.1315	0.1317	0.1318
10	0.1017	0.1040	0.1063	0.1084	0.1104	0.1123	0.1140	0.1157	0.1172	0.1186
11	0.0749	0.0776	0.0802	0.0828	0.0853	0.0878	0.0902	0.0925	0.0948	0.0970
12	0.0505	0.0530	0.0555	0.0579	0.0604	0.0629	0.0654	0.0679	0.0703	0.0728
13	0.0315	0.0334	0.0354	0.0374	0.0395	0.0416	0.0438	0.0459	0.0481	0.0504
14	0.0182	0.0196	0.0210	0.0225	0.0240	0.0256	0.0272	0.0289	0.0306	0.0324
15	0.0098	0.0107	0.0116	0.0126	0.0136	0.0147	0.0158	0.0169	0.0182	0.1094
16	0.0050	0.0055	0.0060	0.0066	0.0072	0.0079	0.0086	0.0093	0.0101	0.0109
17	0.0024	0.0026	0.0029	0.0033	0.0036	0.0040	0.0044	0.0048	0.0053	0.0058
18	0.0011	0.0012	0.0014	0.0015	0.0017	0.0019	0.0021	0.0024	0.0026	0.0029
19	0.0005	0.0005	0.0006	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012	0.0014
20	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004	0.0005	0.0005	0.0006
21	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0002	0.0003
22	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

TABLA 7 PROBABILIDADES POISSON (*continuación*)

	μ									
x	**9.1**	**9.2**	**9.3**	**9.4**	**9.5**	**9.6**	**9.7**	**9.8**	**9.9**	**10**
0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
1	0.0010	0.0009	0.0009	0.0008	0.0007	0.0007	0.0006	0.0005	0.0005	0.0005
2	0.0046	0.0043	0.0040	0.0037	0.0034	0.0031	0.0029	0.0027	0.0025	0.0023
3	0.0140	0.0131	0.0123	0.0115	0.0107	0.0100	0.0093	0.0087	0.0081	0.0076
4	0.0319	0.0302	0.0285	0.0269	0.0254	0.0240	0.0226	0.0213	0.0201	0.0189
5	0.0581	0.0555	0.0530	0.0506	0.0483	0.0460	0.0439	0.0418	0.0398	0.0378
6	0.0881	0.0851	0.0822	0.0793	0.0764	0.0736	0.0709	0.0682	0.0656	0.0631
7	0.1145	0.1118	0.1091	0.1064	0.1037	0.1010	0.0982	0.0955	0.0928	0.0901
8	0.1302	0.1286	0.1269	0.1251	0.1232	0.1212	0.1191	0.1170	0.1148	0.1126
9	0.1317	0.1315	0.1311	0.1306	0.1300	0.1293	0.1284	0.1274	0.1263	0.1251
10	0.1198	0.1210	0.1219	0.1228	0.1235	0.1241	0.1245	0.1249	0.1250	0.1251
11	0.0991	0.1012	0.1031	0.1049	0.1067	0.1083	0.1098	0.1112	0.1125	0.1137
12	0.0752	0.0776	0.0799	0.0822	0.0844	0.0866	0.0888	0.0908	0.0928	0.0948
13	0.0526	0.0549	0.0572	0.0594	0.0617	0.0640	0.0662	0.0685	0.0707	0.0729
14	0.0342	0.0361	0.0380	0.0399	0.0419	0.0439	0.0459	0.0479	0.0500	0.0521
15	0.0208	0.0221	0.0235	0.0250	0.0265	0.0281	0.0297	0.0313	0.0330	0.0347
16	0.0118	0.0127	0.0137	0.0147	0.0157	0.0168	0.0180	0.0192	0.0204	0.0217
17	0.0063	0.0069	0.0075	0.0081	0.0088	0.0095	0.0103	0.0111	0.0119	0.0128
18	0.0032	0.0035	0.0039	0.0042	0.0046	0.0051	0.0055	0.0060	0.0065	0.0071
19	0.0015	0.0017	0.0019	0.0021	0.0023	0.0026	0.0028	0.0031	0.0034	0.0037
20	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012	0.0014	0.0015	0.0017	0.0019
21	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0007	0.0008	0.0009
22	0.0001	0.0001	0.0002	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004
23	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001

	μ									
x	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0010	0.0004	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0037	0.0018	0.0008	0.0004	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
4	0.0102	0.0053	0.0027	0.0013	0.0006	0.0003	0.0001	0.0001	0.0000	0.0000
5	0.0224	0.0127	0.0070	0.0037	0.0019	0.0010	0.0005	0.0002	0.0001	0.0001
6	0.0411	0.0255	0.0152	0.0087	0.0048	0.0026	0.0014	0.0007	0.0004	0.0002
7	0.0646	0.0437	0.0281	0.0174	0.0104	0.0060	0.0034	0.0018	0.0010	0.0005
8	0.0888	0.0655	0.0457	0.0304	0.0194	0.0120	0.0072	0.0042	0.0024	0.0013
9	0.1085	0.0874	0.0661	0.0473	0.0324	0.0213	0.0135	0.0083	0.0050	0.0029
10	0.1194	0.1048	0.0859	0.0663	0.0486	0.0341	0.0230	0.0150	0.0095	0.0058
11	0.1194	0.1144	0.1015	0.0844	0.0663	0.0496	0.0355	0.0245	0.0164	0.0106
12	0.1094	0.1144	0.1099	0.0984	0.0829	0.0661	0.0504	0.0368	0.0259	0.0176
13	0.0926	0.1056	0.1099	0.1060	0.0956	0.0814	0.0658	0.0509	0.0378	0.0271
14	0.0728	0.0905	0.1021	0.1060	0.1024	0.0930	0.0800	0.0655	0.0514	0.0387

TABLA 7 PROBABILIDADES POISSON (*continuación*)

	μ									
x	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
15	0.0534	0.0724	0.0885	0.0989	0.1024	0.0992	0.0906	0.0786	0.0650	0.0516
16	0.0367	0.0543	0.0719	0.0866	0.0960	0.0992	0.0963	0.0884	0.0772	0.0646
17	0.0237	0.0383	0.0550	0.0713	0.0847	0.0934	0.0963	0.0936	0.0863	0.0760
18	0.0145	0.0256	0.0397	0.0554	0.0706	0.0830	0.0909	0.0936	0.0911	0.0844
19	0.0084	0.0161	0.0272	0.0409	0.0557	0.0699	0.0814	0.0887	0.0911	0.0888
20	0.0046	0.0097	0.0177	0.0286	0.0418	0.0559	0.0692	0.0798	0.0866	0.0888
21	0.0024	0.0055	0.0109	0.0191	0.0299	0.0426	0.0560	0.0684	0.0783	0.0846
22	0.0012	0.0030	0.0065	0.0121	0.0204	0.0310	0.0433	0.0560	0.0676	0.0769
23	0.0006	0.0016	0.0037	0.0074	0.0133	0.0216	0.0320	0.0438	0.0559	0.0669
24	0.0003	0.0008	0.0020	0.0043	0.0083	0.0144	0.0226	0.0328	0.0442	0.0557
25	0.0001	0.0004	0.0010	0.0024	0.0050	0.0092	0.0154	0.0237	0.0336	0.0446
26	0.0000	0.0002	0.0005	0.0013	0.0029	0.0057	0.0101	0.0164	0.0246	0.0343
27	0.0000	0.0001	0.0002	0.0007	0.0016	0.0034	0.0063	0.0109	0.0173	0.0254
28	0.0000	0.0000	0.0001	0.0003	0.0009	0.0019	0.0038	0.0070	0.0117	0.0181
29	0.0000	0.0000	0.0001	0.0002	0.0004	0.0011	0.0023	0.0044	0.0077	0.0125
30	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0013	0.0026	0.0049	0.0083
31	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0007	0.0015	0.0030	0.0054
32	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0004	0.0009	0.0018	0.0034
33	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0005	0.0010	0.0020
34	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0012
35	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0007
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004
37	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002
38	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
39	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

TABLA 8 VALORES T_L PARA LA PRUEBA DE MANN-WHITNEY-WILCOXON
Rechazar la hipótesis de que las poblaciones son idénticas si la suma de los rangos de las n_1 elementos es *menor* que el valor T_L de la tabla siguiente o si la suma de los rangos de los n_1 elementos es *mayor* que el valor T_U donde

$$T_U = n_1(n_1 + n_2 + 1) - T_L$$

		n_2								
α = 0.10		**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
n_1	2	3	3	3	4	4	4	5	5	5
	3	6	7	7	8	9	9	10	11	11
	4	10	11	12	13	14	15	16	17	18
	5	16	17	18	20	21	22	24	25	27
	6	22	24	25	27	29	30	32	34	36
	7	29	31	33	35	37	40	42	44	46
	8	38	40	42	45	47	50	52	55	57
	9	47	50	52	55	58	61	64	67	70
	10	57	60	63	67	70	73	76	80	83

		n_2								
α = 0.05		**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
n_1	2	3	3	3	3	3	3	4	4	4
	3	6	6	6	7	8	8	9	9	10
	4	10	10	11	12	13	14	15	15	16
	5	15	16	17	18	19	21	22	23	24
	6	21	23	24	25	27	28	30	32	33
	7	28	30	32	34	35	37	39	41	43
	8	37	39	41	43	45	47	50	52	54
	9	46	48	50	53	56	58	61	63	66
	10	56	59	61	64	67	70	73	76	79

Apéndice C: Notación para la suma

Suma

Definición

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \tag{C.1}$$

Ejemplo para $x_1 = 5$, $x_2 = 8$, $x_3 = 14$:

$$\begin{aligned}\sum_{i=1}^{3} x_i &= x_1 + x_2 + x_3 \\ &= 5 + 8 + 14 \\ &= 27\end{aligned}$$

Resultado 1
Para una constante c:

$$\sum_{i=1}^{n} c = \underbrace{(c + c + \cdots + c)}_{n \text{ veces}} = nc \tag{C.2}$$

Ejemplo para $c = 5$, $n = 10$:

$$\sum_{i=1}^{10} 5 = 10(5) = 50$$

Ejemplo para $c = \bar{x}$:

$$\sum_{i=1}^{n} \bar{x} = n\bar{x}$$

Resultado 2

$$\begin{aligned}\sum_{i=1}^{n} cx_i &= cx_1 + cx_2 + \cdots + cx_n \\ &= c(x_1 + x_2 + \cdots + x_n) = c\sum_{i=1}^{n} x_i\end{aligned} \tag{C.3}$$

Ejemplo para $x_1 = 5$, $x_2 = 8$, $x_3 = 14$, $c = 2$:

$$\sum_{i=1}^{3} 2x_i = 2\sum_{i=1}^{3} x_i = 2(27) = 54$$

Resultado 3

$$\sum_{i=1}^{n} (ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i \tag{C.4}$$

Ejemplo para $x_1 = 5$, $x_2 = 8$, $x_3 = 14$, $a = 2$, $y_1 = 7$, $y_2 = 3$, $y_3 = 8$, $b = 4$:

$$\begin{aligned}\sum_{i=1}^{3}(2x_i + 4y_i) &= 2\sum_{i=1}^{3}x_i + 4\sum_{i=1}^{3}y_i \\ &= 2(27) + 4(18) \\ &= 54 + 72 \\ &= 126\end{aligned}$$

Doble suma

Considere los datos siguientes en los que interviene la variable x_{ij}, donde i es el subíndice que denota la posición en un renglón y j es el subíndice que denota la posición en una columna.

		Columna	
	1	**2**	**3**
Renglón 1	$x_{11} = 10$	$x_{12} = 8$	$x_{13} = 6$
Renglón 2	$x_{21} = 7$	$x_{22} = 4$	$x_{23} = 12$

Definición

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij} = {} & (x_{11} + x_{12} + \cdots + x_{1m}) + (x_{21} + x_{22} + \cdots + x_{2m}) \\ & + (x_{31} + x_{32} + \cdots + x_{3m}) + \cdots + (x_{n1} + x_{n2} + \cdots + x_{nm})\end{aligned} \tag{C.5}$$

Ejemplo:

$$\begin{aligned}\sum_{i=1}^{2}\sum_{j=1}^{3}x_{ij} &= x_{11} + x_{12} + x_{13} + x_{21} + x_{22} + x_{23} \\ &= 10 + 8 + 6 + 7 + 4 + 12 \\ &= 47\end{aligned}$$

Definición

$$\sum_{i=1}^{n}x_{ij} = x_{1j} + x_{2j} + \cdots + x_{nj} \tag{C.6}$$

Ejemplo:

$$\begin{aligned}\sum_{i=1}^{2}x_{i2} &= x_{12} + x_{22} \\ &= 8 + 4 \\ &= 12\end{aligned}$$

Notación abreviada

Algunas veces cuando se suma, sobre todo los valores de un subíndice, se usa la siguiente notación abreviada:

$$\sum_{i=1}^{n}x_i = \sum x_i \tag{C.7}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij} = \sum\sum x_{ij} \tag{C.8}$$

$$\sum_{i=1}^{n}x_{ij} = \sum_{i}x_{ij} \tag{C.9}$$

Apéndice D: Soluciones para los autoexámenes y respuestas a los ejercicios con números pares

Capítulo 1

2. a. 9
b. 4
c. Cualitativas: país y precio de la habitación
Cuantitativas: cantidad de habitaciones y evaluación general.
d. País es nominal; precio de la habitación es ordinal; número de habitaciones es de razón; evaluación general es de intervalo.

3. a. La cantidad promedio de habitaciones = 808/9 = 89.78, o aproximadamente 90 habitaciones
b. El promedio de las puntuaciones = 732.1/19 = 81.3
c. 2 de 9 se encuentran en Inglaterra; aproximadamente 22%
d. En 4 de 9 el precio es $$; aproximadamente 44%

4. a. 10
b. Todas las marcas de minicomponentes
c. $314
d. $314

6. Las preguntas a, c y d proporcionan datos cuantitativos
Las preguntas b y e proporcionan datos cualitativos

8. a. 1 005
b. Cualitativos
c. Porcentajes
d. Aproximadamente 291

10. a. Cuantitativo; de razón
b. Cualitativo; nominal
c. Cualitativo; ordinal
d. Cualitativo; de razón
e. Cualitativo; nominal

12. a. Todas las personas que visitan Hawai
b. Sí
c. Las preguntas primera y cuarta proporcionan datos cuantitativos
Las preguntas segunda y tercera proporcionan datos cualitativos

13. a. Las ganancias en miles de millones de dólares son datos cuantitativos
b. De serie de tiempo de 1997 a 2005
c. Las ganancias de Volkswagen
d. Las ganancias son relativamente bajas de 1997 a 1999, hay un crecimiento excelente en 2000 y 2001 y de 2003 a 2005 hay una disminución; la disminución en las ganancias sugiere que las ganancias de $600 millones proyectadas para 2006 son razonables
e. En julio de 2001, la tendencia en las ganancias era positiva; en el 2001 Volskwagen debe haber sido una inversión prometedora
f. Ser cuidadosos al proyectar datos de series de tiempo hacia el futuro, ya que las tendencias de los datos del pasado pueden continuar o no

14. a. Gráfica con una línea de serie de tiempo para cada fabricante
b. Toyota sobrepasa a General Motors en 2006 y se convierte en el principal fabricantes de automóviles
c. En una gráfica de barras se pueden mostrar datos de sección transversal para el 2007; las alturas de las barras serán GM 8.8, Ford 7.9, DC 4.6 y Toyota 9.6

16. a. Pruebas de sabor del producto y pruebas de marketing
b. Mediante un estudio estadístico diseñado especialmente

18. a. 36%
b. 189
c. Cualitativos

20. a. 43% de los dirigentes se clasificaron a sí mismos como optimistas o muy optimistas y 21% eligieron la atención de la salud como el sector con más probabilidad de ir a la cabeza del mercado en los próximos 12 meses
b. El rendimiento promedio esperado por la población de todos los directivos de inversiones durante los próximos 12 meses es 11.2%
c. El promedio muestral de 2.5 años es una estimación del tiempo que la población de administradores de inversiones cree que se necesitará para que recobren un crecimiento sustancial

22. a. Todos los votantes registrados de California
b. Los votantes registrados contactados para la encuesta
c. Porque se necesita demasiado tiempo y dinero para contactar a toda la población

24. a. Correcto
b. Incorrecto
c. Correcto
d. Incorrecto
e. Incorrecto

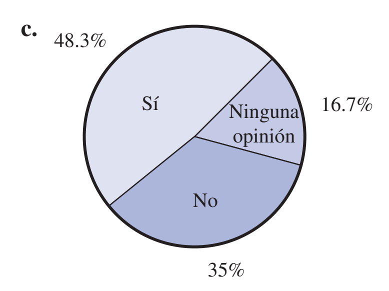
Capítulo 2

2. a. 0.20
b. 40
c/d.

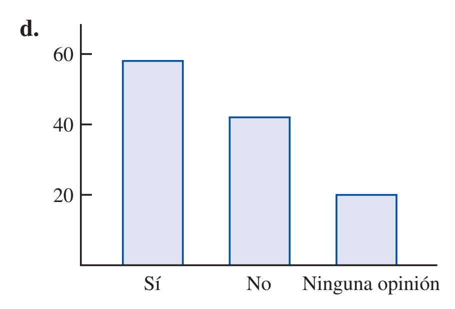
Clase	Frecuencia	Frecuencia porcentual
A	44	22
B	36	18
C	80	40
D	40	20
Total	200	100

3. a. $360° \times 58/120 = 174°$
b. $360° \times 42/120 = 126°$

c.



d.



4. **a.** Cualitativos

b.

Programa de TV	Frecuencia	Frecuencia porcentual
CSI	18	36
ER	11	22
Friends	15	30
Raymond	6	12
Total	50	100

d. *CSI* es el que tiene la mayor audiencia; *Friends* tiene el segundo lugar

6. **a.**

Cadena de TV	Frecuencia	Frecuencia porcentual
ABC	15	30
CBS	17	34
FOX	1	2
NBC	17	34

b. CBS y NBC empatan en el primer lugar; ABC está cerca con 15

7.

Evaluación	Frecuencia	Frecuencia relativa
Óptimo	19	0.38
Muy bueno	13	0.26
Bueno	10	0.20
Regular	6	0.12
Malo	2	0.04

La administración debe estar satisfecha con estos resultados: 64% de las evaluaciones son de muy bueno a óptimo, y 84% de las evaluaciones corresponden a bueno, muy bueno u óptimo; mediante una comparación de estas evaluaciones con resultados previos se podrá ver si el restaurante ha mejorado en las evaluaciones de sus clientes en cuanto a la calidad de los alimentos

8. **a.**

Posición	Frecuencia	Frecuencia relativa
P	17	0.309
H	4	0.073
1	5	0.091
2	4	0.073
3	2	0.036
S	5	0.091
L	6	0.109
C	5	0.091
R	7	0.127
Totales	55	1.000

b. Pitcher
c. 3era. base
d. Right field
e. 16 jugadores dentro del diamante en comparación con 18 jugadores fuera del diamante

10. **a.** Éstos son cualitativos; dan una clasificación cualitativa.

b.

Evaluación	Frecuencia	Frecuencia relativa
1 estrellas	0	0.000
2 estrellas	3	0.167
3 estrellas	3	0.167
4 estrellas	10	0.556
5 estrellas	2	0.111
	18	1.000

d. En general fue muy buena, 10 de las evaluaciones le dieron 4 estrellas y 12 (66.7%) le dieron 4 o 5 estrellas

12.

Clase	Frecuencia acumulada	Frecuencia relativa acumulada
≤19	10	0.20
≤29	24	0.48
≤39	41	0.82
≤49	48	0.96
≤59	50	1.00

14. **b/c.**

Clase	Frecuencia acumulada	Frecuencia porcentual
6.0–7.9	4	20
8.0–9.9	2	10
10.0–11.9	8	40
12.0–13.9	3	15
14.0–15.9	3	15
Totales	20	100

15. a/b.

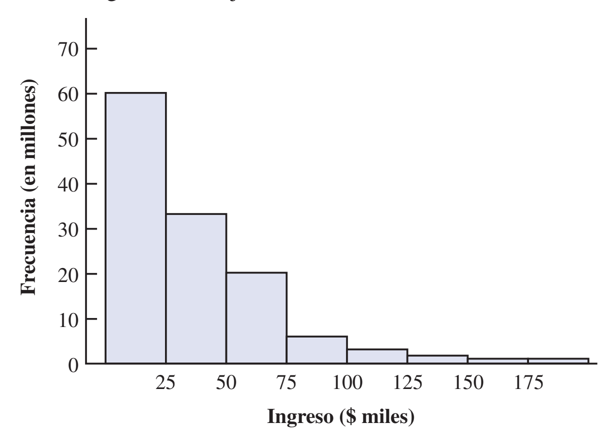
Tiempo de espera	Frecuencia	Frecuencia relativa
0–4	4	0.20
5–9	8	0.40
10–14	5	0.25
15–19	2	0.10
20–24	1	0.05
Totales	20	1.00

c/d.

Tiempo de espera	Frecuencia acumulada	Frecuencia relativa acumulada
≤4	4	0.20
≤9	12	0.60
≤14	17	0.85
≤19	19	0.95
≤24	20	1.00

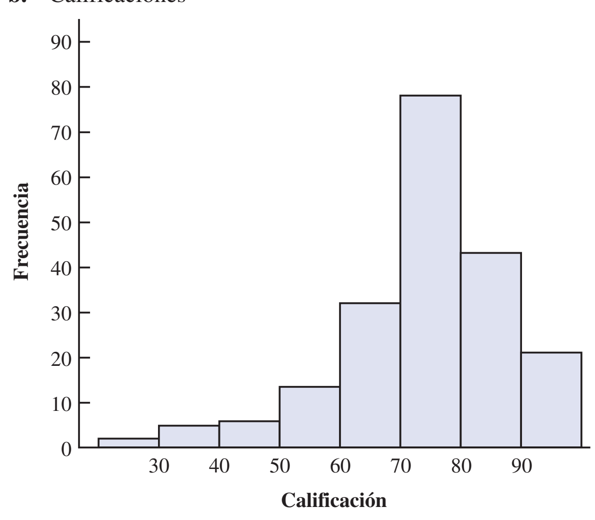
e. 12/20 = 0.60

16. a. Ingreso bruto ajustado



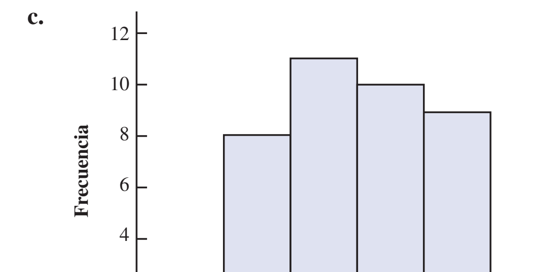
Este histograma está sesgado a la derecha

b. Calificaciones



Este histograma está sesgado a la izquierda

c.



Este histograma está sesgado a la izquierda.

18. a. Menor $180; mayor $2 050

b.

Gasto	Frecuencia	Frecuencia porcentual
$0–249	3	12
250–499	6	24
500–749	5	20
750–999	5	20
1000–1249	3	12
1250–1499	1	4
1500–1749	0	0
1750–1999	1	4
2000–2249	1	4
Total	25	100

c. Esta distribución muestra un sesgo positivo

d. La mayoría de los consumidores (64%) gasta entre $250 y $1 000; el valor intermedio es aproximadamente $750, y dos personas gastaron más de $1 750

20. a.

Precio	Frecuencia	Frecuencia porcentual
30–39.99	7	35
40–49.99	5	25
50–59.99	2	10
60–69.99	3	15
70–79.99	3	15
Total	20	100

c. Fleetwood Mac; Harper/Johnson

22.

5	7 8
6	4 5 8
7	0 2 2 5 5 6 8
8	0 2 3 5

23. Unidad de hoja = 0.1

6	3
7	5 5 7
8	1 3 4 8
9	3 6
10	0 4 5
11	3

24. Unidad de hoja = 10

11	6
12	0 2
13	0 6 7
14	2 2 7
15	5
16	0 2 8
17	0 2 3

25.

9	8 9
10	2 4 6 6
11	4 5 7 8 8 9
12	2 4 5 7
13	1 2
14	4
15	1

26. a.

1	0 3 7 7
2	4 5 5
3	0 0 5 5 9
4	0 0 0 5 5 8
5	0 0 0 4 5 5

b.

0	5 7
1	0 1 1 3 4
1	5 5 5 8
2	0 0 0 0 0 0
2	5 5
3	0 0 0
3	6
4	
4	
5	
5	
6	3

28. a.

2	14
2	67
3	011123
3	5677
4	003333344
4	6679
5	00022
5	5679
6	14
6	6
7	2

b. 40–44 tuvo 9
c. 43 tuvo 5
d. 10%; participación relativamente baja en esta carrera

29. a.

		y 1	2	Total
x	A	5	0	5
	B	11	2	13
	C	2	10	12
	Total	18	12	30

b.

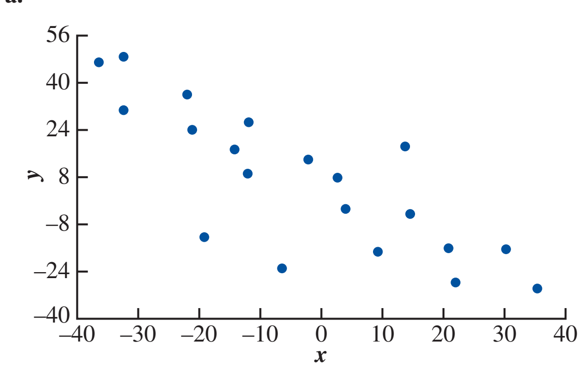
		y 1	2	Total
x	A	100.0	0.0	100.0
	B	84.6	15.4	100.0
	C	16.7	83.3	100.0

c.

		y 1	2
x	A	27.8	0.0
	B	61.1	16.7
	C	11.1	83.3
	Total	100.0	100.0

d. Todos los valores de *A* corresponden a $y = 1$
La mayor parte de los valores de *B* corresponden a $y = 1$
La mayor parte de los valores de *C* corresponden a $y = 2$

30. a.



b. Entre *x* y *y* hay una relación negativa; *y* disminuye a medida que *x* aumenta

32. a.

	Ingreso por familia (en miles de dólares)					
Nivel de estudios	Menos 25	25.0–49.9	50.0–74.9	75.0–99.9	100 o más	Total
No terminó secundaria	32.70	14.82	8.27	5.02	2.53	15.86
Terminó secundaria	35.74	35.56	31.48	25.39	14.47	30.78
Parte del bachillerato	21.17	29.77	30.25	29.82	22.26	26.37
Título universitario	7.53	14.43	20.56	25.03	33.88	17.52
Posgrado	2.86	5.42	9.44	14.74	26.86	9.48
Total	100.00	100.00	100.00	100.00	100.00	100.00

15.86% de los jefes de familia no terminó la secundaria
b. 26.86%, 39.72%
c. Relación positiva entre ingreso y nivel de educación

34. a.

Ventas/ Margen/ ROE	EPS 0–19	20–39	40–59	60–79	80–100	Total
A				1	8	9
B		1	4	5	2	12
C	1		1	2	3	7
D	3	1		1		5
E		2	1			3
Total	4	4	6	9	13	36

b.

Ventas/ Margen/ ROE	EPS 0–19	20–39	40–59	60–79	80–100	Total
A				11.11	88.89	100
B		8.33	33.33	41.67	16.67	100
C	14.29		14.29	28.57	42.86	100
D	60.00	20.00		20.00		100
E		66.67	33.33			100

Evaluaciones EPS más altas parecen estar relacionadas con evaluaciones más altas sobre Ventas/Margen/ROE

36. b. Ninguna relación aparente

38. a.

Vehículo	Frecuencia	Frecuencia porcentual
Accord	6	12
Camry	7	14
F-Series	14	28
Ram	10	20
Silverado	13	26

b. Ford F-Series y Toyota Camry

40. a.

Respuesta	Frecuencia	Frecuencia porcentual
Precisión	16	16
Técnica de golpe	3	3
Actitud mental	17	17
Energía	8	8
Practice	15	15
Práctica	10	10
Tiro al hoyo	24	24
Estrategia de decisión	7	7
Total	100	100

b. Juego corto inadecuado, actitud mental inadecuada, falta de precisión y práctica insuficiente

42. a.

Puntuación en el SAT	Frecuencia
750–849	2
850–949	5
950–1049	10
1050–1149	5
1150–1249	3
Total	25

b. Casi simétrica

c. El 40% de las puntuaciones se encuentra entre 950 y 1 049. Puntuaciones menores que 750 o mayores que 1 249 son poco usuales. La media es encuentra un poco arriba de 1 000.

44. a.

Población	Frecuencia	Frecuencia porcentual
0.0–2.4	17	34
2.5–4.9	12	24
5.0–7.4	9	18
7.5–9.9	4	8
10.0–12.4	3	6
12.5–14.9	1	2
15.0–17.4	1	2
17.5–19.9	1	2
20.0–22.4	0	0
22.5–24.9	1	2
25.0–27.4	0	0
27.5–29.9	0	0
30.0–32.4	0	0
32.5–34.9	0	0
35.0–37.4	1	2
Total	50	100

c. Sesgo ligeramente positivo

d. 17 (34%) tienen una población menor que 2.5 millones
29 (58%) tienen una población menor que 5 millones
8 (16%) tienen una población mayor que 10 millones
El más grande tiene 35.9 millones (California)
El menor 0.5 millones (Wyoming)

46. a. Temperaturas altas

1	
2	
3	0
4	1 2 2 5
5	2 4 5
6	0 0 0 1 2 2 5 6 8
7	0 7
8	4

b. Temperaturas bajas

1	1
2	1 2 6 7 9
3	1 5 6 8 9
4	0 3 3 6 7
5	0 0 4
6	5
7	
8	

c. La mayor frecuencia entre las temperaturas altas se observa en los 60 (9 de 20) y sólo hay una temperatura inferior a 54
La mayor parte de las temperaturas altas están entre 41 y 68, mientras que la mayor parte de las temperaturas bajas se encuentran entre 21 y 47
La más baja fue 11 y la más alta 84

d.

Temperatura alta	Frecuencia	Temperatura baja	Frecuencia
10–19	0	10–19	1
20–29	0	20–29	5
30–39	1	30–39	5
40–49	4	40–49	5
50–59	3	50–59	3
60–69	9	60–69	1
70–79	2	70–79	0
80–89	1	80–89	0
Total	20	Total	20

48. a.

	Satisfacción en el empleo						
Ocupación	**30–39**	**40–49**	**50–59**	**60–69**	**70–79**	**80–89**	**Total**
Ebanista			2	4	3	1	10
Abogado	1	5	2	1	1		10
Terapeuta físico			5	2	1	2	10
Analista de sistemas		2	1	4	3		10
Total	1	7	10	11	8	3	40

b.

	Satisfacción en el empleo						
Ocupación	**30–39**	**40–49**	**50–59**	**60–69**	**70–79**	**80–89**	**Total**
Ebanista			20	40	30	10	100
Abogado	10	50	20	10	10		100
Terapeuta físico			50	20	10	20	100
Analista de sistemas		20	10	40	30		100

c. Los ebanistas parecen ser los que tienen mayor satisfacción en el empleo; los abogados parecen ser los que tienen menor satisfacción en el empleo

50. a. Totales de los renglones: 247; 54; 82; 121
Totales de las columnas: 149; 317; 17; 7; 14

b.

Año	Frec.	Combustible	Frec.
1973 o antes	247	Elect.	149
1974–79	54	Petróleo	317
1980–86	82	Oil	17
1987–91	121	Propano	7
Total	504	Otros	14
		Total	504

c. Tabulación cruzada con los porcentajes de las columnas

	Tipo de combustible				
Año de construcción	**Electricidad**	**Gas natural**	**Petróleo**	**Propano**	**Otros**
1973 o antes	26.9	57.7	70.5	71.4	50.0
1974–1979	16.1	8.2	11.8	28.6	0.0
1980–1986	24.8	12.0	5.9	0.0	42.9
1987–1991	32.2	22.1	11.8	0.0	7.1
Total	100.0	100.0	100.0	100.0	100.0

d. Tabulación cruzada con los porcentajes de los renglones

	Tipo de combustible					
Año de construcción	**Electricidad**	**Gas natural**	**Petróleo**	**Propano**	**Otros**	**Total**
1973 o antes	16.2	74.1	4.9	2.0	2.8	100.0
1974–1979	44.5	48.1	3.7	3.7	0.0	100.0
1980–1986	45.1	46.4	1.2	0.0	7.3	100.0
1987–1991	39.7	57.8	1.7	0.0	0.8	100.0

52. a. Tabulación cruzada de valor de mercado y ganancia

	Valor de mercado ($ miles)				
Ganancias ($ miles)	**0–300**	**300–600**	**600–900**	**900–1200**	**Total**
0–8000	23	4			27
8000–16 000	4	4	2	2	12
16 000–24 000		2	1	1	4
24 000–32 000		1	2	1	4
32 000–40 000		2	1		3
Total	27	13	6	4	50

b. Tabulación cruzada de los porcentajes de renglón

	Ganancias ($ miles)				
Valor de mercado ($ miles)	**0–300**	**300–600**	**600–900**	**900–1200**	**Total**
0–8000	85.19	14.81	0.00	0.00	100
8000–16 000	33.33	33.33	16.67	16.67	100
16 000–24 000	0.00	50.00	25.00	25.00	100
24 000–32 000	0.00	25.00	50.00	25.00	100
32 000–40 000	0.00	66.67	33.33	0.00	100

c. Parece haber una relación positiva entre ganancias y valor de mercado; a medida que las ganancias aumentan, aumenta el valor de mercado

54. b. Se demuestra que existe una relación positiva entre valor de mercado y fondos propios

Capítulo 3

2. 16, 16.5

3. Se ordenan los datos de menor a mayor: 15, 20, 25, 25, 27, 28, 30, 34 $i = \frac{20}{100}(8) = 1.6$; redondear hacia arriba a la posición 2
percentil 20 $= 20$

$i = \frac{25}{100}(8) = 2$; usar las posiciones 2 y 3

percentil 25 $= \frac{20 + 25}{2} = 22.5$

$i = \frac{65}{100}(8) = 5.2$; redondear hacia arriba a la posición 6

percentil 65 $= 28$

$i = \frac{75}{100}(8) = 6$; usar las posiciones 6 y 7

percentil 75 $= \frac{28 + 30}{2} = 29$

4. 59.73, 57, 53

6. a. Marketing: 36.3, 35.5, 34.2
Contaduría: 45.7, 44.7, no hay moda

b. Marketing: 34.2, 39.5
Contaduría: 40.95, 49.8

c. Los salarios de los contadores son mayores en aproximadamente $9 000

8. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{695}{20} = 34.75$

Moda = 25 (aparece tres veces)

b. Datos ordenados de menor a mayor: 18, 20, 25, 25, 25, 26, 27, 27, 28, 33, 36, 37, 40, 40, 42, 45, 46, 48, 53, 54

Mediana (posiciones 10 y 11)

$$\frac{33 + 36}{2} = 34.5$$

Las personas que trabajan desde su hogar son un poco más jóvenes.

c. $i = \frac{25}{100}(20) = 5$; usar posiciones 5 y 6

$Q_1 = \frac{25 + 26}{2} = 25.5$

$i = \frac{75}{100}(20) = 15$; usar las posiciones 15 y 16

$Q_3 = \frac{42 + 45}{2} = 43.5$

d. $i = \frac{32}{100}(20) = 6.4$; redondear hacia arriba a la posición 7

percentil 32 = 27

Por lo menos 32% de las personas tienen 27 años o menos

10. a. 76, 76

b. 39, 37.5

c. Sí; los tiempos de espera para las salas de emergencia son muy grandes

12. Disney: 3321, 255.5, 253, 169, 325
Pixar: 3231, 538.5, 505, 363, 631
Las películas de Pixar generan aproximadamente el doble de ganancias por película

14. 16, 4

15. Rango = 34 − 15 = 19
Ordenar los datos de menor a mayor: 15, 20, 25, 25, 27, 28, 30, 34

$i = \frac{25}{100}(8) = 2;\ Q_1 = \frac{20 + 25}{2} = 22.5$

$i = \frac{75}{100}(8) = 6;\ Q_3 = \frac{28 + 30}{2} = 29$

$\text{IQR} = Q_3 - Q_1 = 29 - 22.5 = 6.5$

$\bar{x} = \frac{\Sigma x_i}{n} = \frac{204}{8} = 25.5$

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
27	1.5	2.25
25	−.5	0.25
20	−5.5	30.25
15	−10.5	110.25
30	4.5	20.25
34	8.5	72.25
28	2.5	6.25
25	−.5	0.25
		242.00

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{242}{8 - 1} = 34.57$$

$$s = \sqrt{34.57} = 5.88$$

16. a. Rango = 190 − 168 = 22

b. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{1068}{6} = 178$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{4^2 + (-10)^2 + 6^2 + 12^2 + (-8)^2 + (-4)^2}{6 - 1} = \frac{376}{5} = 75.2$$

c. $s = \sqrt{75.2} = 8.67$

d. $\frac{s}{\bar{x}}(100) = \frac{8.67}{178}(100\%) = 4.87\%$

18. a. 38, 97, 9.85

b. En el este se observa mayor variación

20. *Dawson:* rango = 2, $s = 0.67$
Clark: rango = 8, $s = 2.58$

22. a. 45.05, 23.98; 57.50, 11.475

b. 190.67, 13.81; 140.63, 11.86

c. 38.02%; 57.97%

d. La variabilidad es mayor en las transacciones con ayuda de un corredor

24. *Tiempos en un cuarto de milla:* $s = 0.0564$, Coef. de Var. = 5.8%

Tiempos en una milla: $s = 0.01295$, Coef. de Var. = 2.9%

26. 0.20, 1.50, 0, −0.50, −2.20

27. Teorema de Chebyshev: *por lo menos* $(1 - 1/z^2)$

a. $z = \frac{40 - 30}{5} = 2;\ 1 - \frac{1}{(2)^2} = 0.75$

b. $z = \frac{45 - 30}{5} = 3;\ 1 - \frac{1}{(3)^2} = 0.89$

c. $z = \frac{38 - 30}{5} = 1.6;\ 1 - \frac{1}{(1.6)^2} = 0.61$

d. $z = \frac{42 - 30}{5} = 2.4;\ 1 - \frac{1}{(2.4)^2} = 0.83$

e. $z = \frac{48 - 30}{5} = 3.6; 1 - \frac{1}{(3.6)^2} = 0.92$

28. a. 95%
b. Casi todos
c. 68%

29. a. $z = 2$ desviaciones estándar

$1 - \frac{1}{z^2} = 1 - \frac{1}{2^2} = \frac{3}{4}$; por lo menos 75%

b. $z = 2.5$ desviaciones estándar

$1 - \frac{1}{z^2} = 1 - \frac{1}{2.5^2} = 0.84$; por lo menos 84%

c. $z = 2$ desviaciones estándar
Regla empírica: 95%

30. a. 68%
b. 81.5%
c. 2.5%

32. a. -0.67
b. 1.50
c. Ninguno es una observación atípica
d. Sí; $z = 8.25$

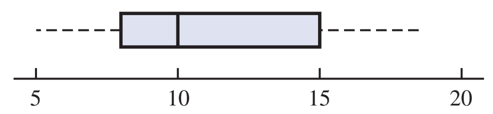
38. Ordenar los datos de menor a mayor: 5, 6, 8, 10, 10, 12, 15, 16, 18

$i = \frac{25}{100}(9) = 2.25$; redondear hacia arriba a la posición 3

$Q_1 = 8$
Mediana (posición 5) = 10

$i = \frac{75}{100}(9) = 6.75$; redondear hacia arriba a la posición 7

$Q_3 = 15$
Resumen de cinco números: 5, 8, 10, 15, 18



40. a. 619, 725, 1 016, 1 699, 4 450
b. Límites: 0, 3 160
c. Sí
d. No

41. a. Ordenar los datos de menor a mayor

$i = \frac{25}{100}(21) = 5.25$; redondear hacia arriba a la posición 6

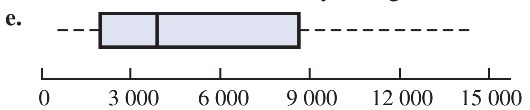
$Q_1 = 1\,872$
Mediana (posición 11) = 4 019

$i = \frac{75}{100}(21) = 15.75$; redondear hacia arriba a la posición 16

$Q_3 = 8\,305$
Resumen de cinco números: 608, 1 872, 4 019, 8 305, 14 138

b. RIC = $Q_3 - Q_1$ = 8 305 − 1 872 = 6 433
Límite inferior: 1 872 − 1.5(6 433) = −7 777.5
Límite superior: 8 305 + 1.5(6 433) = 17 955

c. No; los datos están dentro de los límites
d. 41 138 > 27 604; 41 138 será un dato atípico; el valor de este dato deberá ser revisado y corregido
e.



42. a. 66
b. 30, 49, 66, 88, 208
c. Sí; el límite superior = 146.5

44. a. 18.2, 15.35
b. 11.7, 23.5
c. 3.4, 11.7, 15.35, 23.5, 41.3
d. Sí; Alger Small Cap 41.3

45. b. Entre las variables x y y parece haber una relación lineal negativa
c.

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
4	50	−4	4	−16
6	50	−2	4	−8
11	40	3	−6	−18
3	60	−5	14	−70
16	30	8	−16	−128
40	230	0	0	−240

$\bar{x} = 8; \bar{y} = 46$

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{-240}{4} = -60$$

La covarianza muestral indica una relación lineal negativa entre x y y

d. $r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-60}{(5.43)(11.40)} = -0.969$

El coeficiente de correlación muestral que es –0.969 indica una fuerte relación lineal negativa

46. b. Parece haber una relación lineal positiva entre x y y
c. $s_{xy} = 26.5$
d. $r_{xy} = 0.693$

48. –0.91; relación negativa

50. b. 0.9098
c. Relación lineal positiva fuerte; no

52. a. 3.69
b. 3.175

53. a

f_i	M_i	$f_i M_i$
4	5	20
7	10	70
9	15	135
5	20	100
25		325

$$\bar{x} = \frac{\Sigma f_i M_i}{n} = \frac{325}{25} = 13$$

b.

f_i	M_i	$(M_i - \bar{x})$	$(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
4	5	−8	64	256
7	10	−3	9	63
9	15	2	4	36
5	20	7	49	245
25				600

$$s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} = \frac{600}{25 - 1} = 25$$

$$s = \sqrt{25} = 5$$

54. a.

Calificación x_i	Peso w_i
4 (A)	9
3 (B)	15
2 (C)	33
1 (D)	3
0 (F)	0
	60 horas crédito

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} = \frac{9(4) + 15(3) + 33(2) + 3(1)}{9 + 15 + 33 + 3}$$

$$= \frac{150}{60} = 2.5$$

b. Sí

56. 3.49, .94

58. a. 1 800, 1 351
b. 387, 1 710
c. 7 280, 1 323
d. 3 675 303, 1 917
e. 9 271.01, 96.29
f. Fuerte sesgo positivo
g. Con un diagrama de caja, 4 135 y 7 450 son observaciones atípicas

60. a. 2.3, 1.85
b. 1.90, 1.38
c. Altria Group 5%
d. −0.51, menor que la media
e. 1.02, mayor que la media
f. No

62. a. $670
b. $456
c. $z = 3$; sí
d. Ahorro de tiempo y evitar el costo de una multa

64. a. 215.9
b. 55%
c. 175.0, 628.3
d. 48.8, 175.0, 215.9, 628.3, 2 325.0
e. Sí, todo precio superior a 1 308.25
f. 482.1; preferiere la mediana

66. b. 0.9856, una fuerte relación positiva

68. a. 817
b. 833

70. a. 60.68
b. $s^2 = 31.23$; $s = 5.59$

Capítulo 4

2. $\binom{6}{3} = \frac{6!}{3!3!} = \frac{6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1}{(3\cdot 2\cdot 1)(3\cdot 2\cdot 1)} = 20$

ABC	ACE	BCD	BEF
ABD	ACF	BCE	CDE
ABE	ADE	BCF	CDF
ABF	ADF	BDE	CEF
ACD	AEF	BDF	DEF

4. b. (H,H,H), (H,H,T), (H,T,H), (H,T,T), (T,H,H), (T,H,T), (T,T,H), (T,T,T)
c. ⅛

6. $P(E_1) = 0.40$, $P(E_2) = 0.26$, $P(E_3) = 0.34$
Se usó el método de la frecuencia relativa

8. a. 4: Comisión positiva—Consejo aprueba
Comisión positiva—Consejo desaprueba
Comisión negativa—Consejo desaprueba
Comisión negativa—Consejo desaprueba

9. $\binom{50}{4} = \frac{50!}{4!46!} = \frac{50\ 49\ 48\ 47}{4\cdot 3\cdot 2\cdot 1} = 230\ 300$

10. a. Con el método de la frecuencia relativa
P(California) = 1 434/2 374 = 0.60
b. Cantidad que no es de ninguno de los cuatro estados
= 2 374 − 1 434 − 390 − 217 − 112
= 221
P(Ninguno de los cuatro estados) = 221/2 374 = 0.09
c. P(No en etapas iniciales) = 1 − 0.22 = 0.78
d. Estimación de la cantidad de empresas de Massachusetts en etapas iniciales de desarrollo = (0.22)390 ≈ 86
e. Si se supone que las cantidades otorgadas no difieren de acuerdo con los estados, se puede multiplicar la probabilidad de que una cantidad sea otorgada a Colorado por el total del capital de riesgo para obtener una estimación.

Estimación de la cantidad destinada a Colorado = (112/2 374)($32.4)
= $1.53 miles de millones

Nota del autor: La cantidad real otorgada a Colorado fue $1.74 miles de millones

12. a. 3 478 761
b. 1/3 478 761
c. 1/146 107 962

14. a. ¼
b. ½
c. ¾

15. a. S = {as de tréboles, as de diamantes, as de corazones, as de espadas}
b. S = {2 de tréboles, 3 de tréboles, ..., 10 de tréboles J de tréboles, Q de tréboles, K de tréboles, A de tréboles}

c. Hay 12; sota, reina o rey con cada uno de los cuatro palos
d. Para a: 4/52 = 1/13 = 0.08
Para b: 13/52 = 1/14 = 0.25
Para c: 12/52 = 0.23

16. a. 36
c. $1/6$
d. $5/18$
e. No; $P(\text{impar}) = P(\text{par}) = \frac{1}{2}$
f. Clásica

17. a. (4, 6), (4, 7), (4, 8)
b. 0.05 + 0.10 + 0.15 = 0.30
c. (2, 8), (3, 8), (4, 8)
d. 0.05 + 0.05 + 0.15 = 0.25
e. 0.15

18. a. $P(0) = 0.05$
b. $P(4 \text{ o } 5) = 0.20$
c. $P(0, 1, \text{ o } 2) = 0.55$

20. a. 0.108
b. 0.096
c. 0.434

22. a. 0.40, 0.40, 0.60
b. 0.80, sí
c. $A^c = \{E_3, E_4, E_5\}$; $C^c = \{E_1, E_4\}$;
$P(A^c) = 0.60$; $P(C^c) = 0.40$
d. (E_1, E_2, E_5); 0.60
e. 0.80

23. a. $P(A) = P(E_1) + P(E_4) + P(E_6)$
$= 0.05 + 0.25 + 0.10 = 0.40$
$P(B) = P(E_2) + P(E_4) + P(E_7)$
$= 0.20 + 0.25 + 0.05 = 0.50$
$P(C) = P(E_2) + P(E_3) + P(E_5) + P(E_7)$
$= 0.20 + 0.20 + 0.15 + 0.05 = 0.60$
b. $A \cup B = \{E_1, E_2, E_4, E_6, E_7\}$;
$P(A \cup B) = P(E_1) + P(E_2) + P(E_4) + P(E_6) + P(E_7)$
$= 0.05 + 0.20 + 0.25 + 0.10 + 0.05$
$= 0.65$
c. $A \cap B = \{E_4\}$; $P(A \cap B) = P(E_4) = 0.25$
d. Sí, son mutuamente excluyentes
e. $B^c = \{E_1, E_3, E_5, E_6\}$;
$P(B^c) = P(E_1) + P(E_3) + P(E_5) + P(E_6)$
$= 0.05 + 0.20 + 0.15 + 0.10$
$= 0.50$

24. a. 0.05
b. 0.70

26. a. 0.30, 0.23
b. 0.17
c. 0.64

28. Sea B = un automóvil rentado por razones de trabajo
P = un automóvil rentado por razones personales
a. $P(B \cup P) = P(B) + P(P) - P(B \cap P)$
$= 0.540 + 0.458 - 0.300$
$= 0.698$
b. $P(\text{por ninguna de las dos}) = 1 - 0.698 = 0.302$

30. a. $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.40}{0.60} = 0.6667$
b. $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{0.40}{0.50} = 0.80$
c. No, porque $P(A \mid B) \neq P(A)$

32. a.

	Sí	No	Total
18 a 34	0.375	0.085	0.46
35 o mayor	0.475	0.065	0.54
Total	0.850	0.150	1.00

b. 46% 18 a 34; 54% 35 y mayores
c. 0.15
d. 0.1848
e. 0.1204
f. 0.5677
g. Mayor probabilidad de No de 18 a 34

33. a.

Razones de su elección

	Calidad	Costo/Conveniencia	Otras	Total
Tiempo completo	0.218	0.204	0.039	0.461
Medio tiempo	0.208	0.307	0.024	0.539
Total	0.426	0.511	0.063	1.000

b. Lo más común es que un estudiante dé el costo o la conveniencia como la primera razón (probabilidad 0.511), la segunda razón es la calidad (probabilidad 0.426)
c. $P(\text{calidad} \mid \text{tiempo completo}) = 0.218/0.461 = 0.473$
d. $P(\text{calidad} \mid \text{medio tiempo}) = 0.208/0.539 = 0.196$
e. Por independencia, se tiene $P(A)P(B) = P(A \cap B)$; de la tabla
$P(A \cap B) = 0.218$, $P(A) = 0.461$, $P(B) = 0.426$
$P(A)P(B) = (0.461)(0.426) = 0.196$
Como $P(A)P(B) \neq P(A \cap B)$, los eventos no son independientes

34. a. 0.44
b. 0.15
c. 0.136
d. 0.106
e. 0.0225
f. 0.0025

36. a. 0.7921
b. 0.9879
c. 0.0121
d. 0.3364, 0.8236, 0.1764
No cometer falta contra Reggie Miller

38. a. 0.70
b. 0.30
c. 0.67, 0.33

d. 0.20, 0.10
e. 0.40
f. 0.20
g. No; $P(S \mid M) \neq P(S)$

39. a. Sí, porque $P(A_1 \cap A_2) = 0$
b. $P(A_1 \cap B) = P(A_1)P(B \mid A_1) = 0.40(.20) = 0.08$
$P(A_2 \cap B) = P(A_2)P(B \mid A_2) = 0.60(.05) = 0.03$
c. $P(B) = P(A_1 \cap B) + P(A_2 \cap B) = 0.08 + 0.03 = 0.11$
d. $P(A_1 \mid B) = \dfrac{0.08}{0.11} = 0.7273$
$P(A_2 \mid B) = \dfrac{0.03}{0.11} = 0.2727$

40. a. 0.10, 0.20, 0.09
b. 0.51
c. 0.26, 0.51, 0.23

42. M = no hacer un pago
D_1 = el cliente deja de cumplir con los pagos
D_2 = el cliente no deja de cumplir con los pagos
$P(D_1) = 0.05$, $P(D_2) = 0.95$, $P(M \mid D_2) = 0.2$,
$P(M \mid D_1) = 1$

a.
$$P(D_1 \mid M) = \frac{P(D_1)P(M \mid D_1)}{P(D_1)P(M \mid D_1) + P(D_2)P(M \mid D_2)} = \frac{(0.05)(1)}{(0.05)(1) + (0.95)(0.2)} = \frac{0.05}{0.24} = 0.21$$
b. Sí, la probabilidad de que el cliente deje de cumplir es mayor que 0.20

44. a. 0.47, 0.53, 0.50, 0.45
b. 0.4963
c. 0.4463
d. 47%, 53%

46. a. 0.68
b. 52
c. 10

48. a. 315
b. 0.29
c. No
d. Republicanos

50. a. 0.76
b. 0.24

54. a. 0.49
b. 0.44
c. 0.54
d. No
e. Sí

56. a. 0.25
b. 0.125
c. 0.0125
d. 0.10
e. No

58. 3.44%

60. a. 0.40
b. 0.67

Capítulo 5

1. a. Cara, Cara (H, H)
Cara, Cruz (H, T)
Cruz, Cara (T, H)
Cruz, Cruz (T, T)
b. x = cantidad de caras en dos lanzamientos
c.

Resultado	Valor de x
(H, H)	2
(H, T)	1
(T, H)	1
(T, T)	0

d. Discreta; puede tomar tres valores: 0, 1 y 2

2. a. x = tiempo en minutos requerido para armar el producto
b. Cualquier valor positivo: $x > 0$
c. Continua

3. Sea Y = oferta de trabajo
N = ninguna oferta
a. $S = \{(Y, Y, Y), (Y, Y, N), (Y, N, Y), (Y, N, N), (N, Y, Y), (N, Y, N), (N, N, Y), (N, N, N)\}$
b. Sea N = número de ofertas de trabajo; N es una variable aleatoria discreta
c.

Resultado experimental	(Y, Y, Y)	(Y, Y, N)	(Y, N, Y)	(Y, N, N)	(N, Y, Y)	(N, Y, N)	(N, N, Y)	(N, N, N)
Valor de N	3	2	2	1	2	1	1	0

4. $x = 0, 1, 2, \ldots, 12$

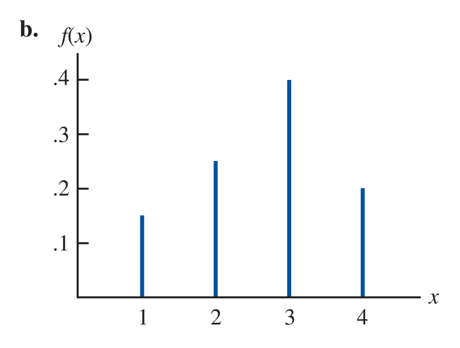
6. a. 0, 1, 2, . . . , 20; discreta
b. 0, 1, 2, . . . ; discreta
c. 0, 1, 2, . . . , 50; discreta
d. $0 \leq x \leq 8$; continua
e. $x > 0$; continua

7. a. $f(x) \geq 0$ para todos los valores de x
$\Sigma f(x) = 1$; por tanto, es una distribución de probabilidad válida
b. Probabilidad de que $x = 30$ es $f(30) = 0.25$
c. Probabilidad de que $x \leq 25$ es
$f(20) + f(25) = 0.20 + 0.15 = 0.35$
d. Probabilidad de que $x > 30$ es $f(35) = 40$

8. a.

x	$f(x)$
1	3/20 = 0.15
2	5/20 = 0.25
3	8/20 = 0.40
4	4/20 = 0.20
	Total 1.00

b.



c. $f(x) \geq 0$ para $x = 1, 2, 3, 4$
$\Sigma f(x) = 1$

10. a.

x	1	2	3	4	5
$f(x)$	0.05	0.09	0.03	0.42	0.41

b.

x	1	2	3	4	5
$f(x)$	0.04	0.10	0.12	0.46	0.28

c. 0.83
d. 0.28
e. Los directivos de alto nivel están más satisfechos con el trabajo

12. a. Sí
b. 0.65

14. a. 0.05
b. 0.70
c. 0.40

16. a.

y	$f(y)$	$yf(y)$
2	0.20	0.4
4	0.30	1.2
7	0.40	2.8
8	0.10	0.8
Totales	1.00	5.2

$E(y) = \mu = 5.2$

b.

y	$y - \mu$	$(y - \mu)^2$	$f(y)$	$(y - \mu)^2 f(y)$
2	−3.20	10.24	0.20	2.048
4	−1.20	1.44	0.30	0.432
7	1.80	3.24	0.40	1.296
8	2.80	7.84	0.10	0.784
			Total	4.560

$\text{Var}(y) = 4.56$

$\sigma = \sqrt{4.56} = 2.14$

18. a/b.

x	$f(x)$	$xf(x)$	$x - \mu$	$(x - \mu)^2$	$(x - \mu)^2 f(x)$
0	0.04	0.00	−1.84	3.39	0.12
1	0.34	0.34	−0.84	0.71	0.24
2	0.41	0.82	0.16	0.02	0.01
3	0.18	0.53	1.16	1.34	0.24
4	0.04	0.15	2.16	4.66	0.17
Total	1.00	1.84			0.79
		↑ $E(x)$			↑ Var(x)

c/d.

y	$f(y)$	$yf(y)$	$y - \mu$	$(y - \mu)^2$	$y - \mu^2 f(y)$
0	0.00	0.00	−2.93	8.58	0.01
1	0.03	0.03	−1.93	3.72	0.12
2	0.23	0.45	−0.93	0.86	0.20
3	0.52	1.55	0.07	0.01	0.00
4	0.22	0.90	1.07	1.15	0.26
Total	1.00	2.93			0.59
		↑ $E(y)$			↑ Var(y)

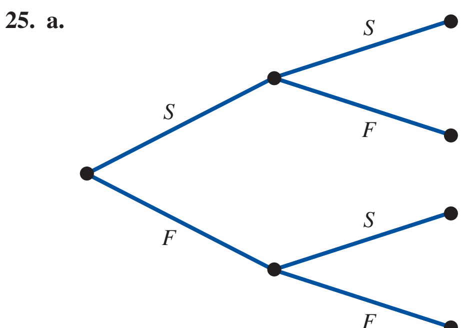
e. El número de recámaras en casas propias es mayor que en casas rentadas; el número esperado de recámaras es 2.93 − 1.84 = 1.09 mayor y la variabilidad en el número de recámaras es menor en las casas propias

20. a. 430
b. −90; la idea es proteger contra los gastos de un accidente grande

22. a. 445
b. Perderá $1 250

24. a. Mediana: 145; grande: 140
b. Mediana: 2 725; grande: 12 400

25. a.



b. $f(1) = \binom{2}{1}(0.4)^1(0.6)^1 = \frac{2!}{1!1!}(0.4)(0.6) = 0.48$

c. $f(0) = \binom{2}{0}(0.4)^0(0.6)^2 = \frac{2!}{0!2!}(1)(0.36) = 0.36$

d. $f(2) = \binom{2}{2}(0.4)^2(0.6)^0 = \frac{2!}{2!0!}(0.16)(0.1) = 0.16$

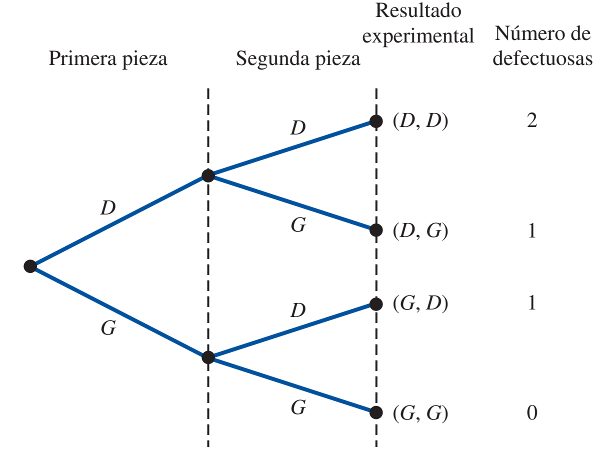
e. $P(x \geq 1) = f(1) + f(2) = 0.48 + 0.16 = 0.64$

f. $E(x) = np = 2(0.4) = 0.8$
$\text{Var}(x) = np(1 - p) = 2(0.4)(0.6) = 0.48$
$\sigma = \sqrt{0.48} = 0.6928$

26. a. $f(0) = 0.3487$
b. $f(2) = 0.1937$
c. 0.9298
d. 0.6513
e. 1
f. $\sigma^2 = 0.9000$, $\sigma = 0.9487$

28. a. 0.2789
b. 0.4181
c. 0.0733

30. a. La probabilidad de que una pieza producida esté defectuosa debe ser 0.03 para toda pieza que se seleccione; las piezas deben seleccionarse independientemente
b. Sea D = defectuosa
G = no defectuosa



c. En dos resultados experimentales hay exactamente una pieza defectuosa
d. $P(\text{ninguna defectuosa}) = (0.97)(0.97) = 0.9409$
$P(\text{1 defectuosa}) = 2(0.03)(0.97) = 0.0582$
$P(\text{2 defectuosas}) = (0.03)(0.03) = 0.0009$

32. a. 0.90
b. 0.99
c. 0.999
d. Sí

34. a. 0.2262
b. 0.8355

38. a. $f(x) = \dfrac{3^x e^{-3}}{x!}$
b. 0.2241
c. 0.1494
d. 0.8008

39. a. $f(x) = \dfrac{2^x e^{-2}}{x!}$
b. $\mu = 6$ en tres lapsos
c. $f(x) = \dfrac{6^x e^{-6}}{x!}$
d. $f(2) = \dfrac{2^2 e^{-2}}{2!} = \dfrac{4(0.1353)}{2} = 0.2706$
e. $f(6) = \dfrac{6^6 e^{-6}}{6!} = 0.1606$
f. $f(5) = \dfrac{4^5 e^{-4}}{5!} = 0.1563$

40. a. $\mu = 48(5/60) = 4$
$f(3) = \dfrac{4^3 e^{-4}}{3!} = \dfrac{(64)(0.0183)}{6} = 0.1952$
b. $\mu = 48(15/60) = 12$
$f(10) = \dfrac{12^{10} e^{-12}}{10!} = 0.1048$
c. $\mu = 48(5/60) = 4$; después de 5 minutos habrá 4 llamadas en espera
$f(0) = \dfrac{4^0 e^{-4}}{0!} = 0.0183$; la probabilidad de que no haya ninguna llamada en espera después de 5 minutos es 0.0183
d. $\mu = 48(3/60) = 2.4$
$f(0) = \dfrac{2.4^0 e^{-2.4}}{0!} = 0.0907$; la probabilidad de que no haya ninguna interrupción en 3 minutos es 0.0907

42. a. $f(0) = \dfrac{7^0 e^{-7}}{0!} = e^{-7} = 0.0009$
b. probabilidad $= 1 - [f(0) + f(1)]$
$f(1) = \dfrac{7^1 e^{-7}}{1!} = 7e^{-7} = 0.0064$
probabilidad $= 1 - [0.0009 + 0.0064] = 0.9927$
c. $\mu = 3.5$
$f(0) = \dfrac{3.5^0 e^{-3.5}}{0!} = e^{-3.5} = 0.0302$
probabilidad $= 1 - f(0) = 1 - 0.0302 = 0.9698$
d.
$$\begin{aligned}\text{probabilidad} &= 1 - [f(0) + f(1) + f(2) + f(3) + f(4)]\\ &= 1 - [0.0009 + 0.0064 + 0.0223 + 0.0521 + 0.0912]\\ &= 0.8271\end{aligned}$$

44. a. $\mu = 1.25$
b. 0.2865
c. 0.3581
d. 0.3554

46. a. $f(1) = \dfrac{\binom{3}{1}\binom{10-3}{4-1}}{\binom{10}{4}} = \dfrac{\left(\frac{3!}{1!2!}\right)\left(\frac{7!}{3!4!}\right)}{\frac{10!}{4!6!}}$
$= \dfrac{(3)(35)}{210} = 0.50$
b. $f(2) = \dfrac{\binom{3}{2}\binom{10-3}{2-2}}{\binom{10}{2}} = \dfrac{(3)(1)}{45} = 0.067$
c. $f(0) = \dfrac{\binom{3}{0}\binom{10-3}{2-0}}{\binom{10}{2}} = \dfrac{(1)(21)}{45} = 0.4667$

d. $f(2) = \dfrac{\binom{3}{2}\binom{10-3}{4-2}}{\binom{10}{4}} = \dfrac{(3)(21)}{210} = 0.30$

48. a. 0.5250
b. 0.1833

50. $N = 60, n = 10$
a. $r = 20, x = 0$

$$f(0) = \frac{\binom{20}{0}\binom{40}{10}}{\binom{60}{10}} = \frac{(1)\left(\frac{40!}{10!30!}\right)}{\frac{60!}{10!50!}}$$
$$= \left(\frac{40!}{10!30!}\right)\left(\frac{10!50!}{60!}\right)$$
$$= \frac{40\cdot39\cdot38\cdot37\cdot36\cdot35\cdot34\cdot33\cdot32\cdot31}{60\cdot59\cdot58\cdot57\cdot56\cdot55\cdot54\cdot53\cdot52\cdot51}$$
$$\approx 0.01$$

b. $r = 20, x = 1$

$$f(1) = \frac{\binom{20}{1}\binom{40}{9}}{\binom{60}{10}} = 20\left(\frac{40!}{9!31!}\right)\left(\frac{10!50!}{60!}\right)$$
$$\approx 0.07$$

c. $1 - f(0) - f(1) = 1 - 0.08 = 0.92$
d. La misma que la probabilidad de que uno sea de Hawai; en el inciso b esta probabilidad fue igual a 0.07

52. a. 0.5333
b. 0.6667
c. 0.7778
d. $n = 7$

54. a.

x	1	2	3	4	5
$f(x)$	0.24	0.21	0.10	0.21	0.24

b. 3.00, 2.34
c. Renta fija: $E(x) = 1.36$, $\text{Var}(x) = 0.23$
Acciones: $E(x) = 4$, $\text{Var}(x) = 1$

56. a. 0.0596
b. 0.3585
c. 100
d. 9.75

58. a. 0.9510
b. 0.0480
c. 0.0490

60. a. 240
b. 12.96
c. 12.96

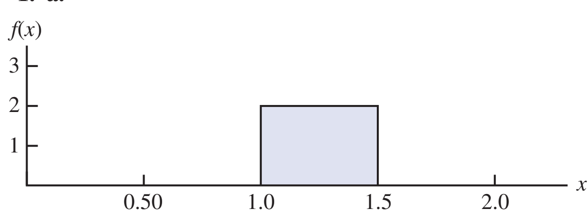
62. 0.1912

64. a. 0.2240
b. 0.5767

66. a. 0.4667
b. 0.4667
c. 0.0667

Capítulo 6

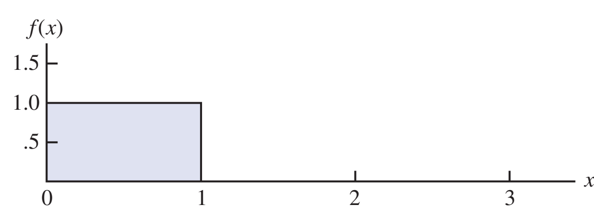
1. a.



b. $P(x = 1.25) = 0$; la probabilidad de cualquier punto es cero porque el área bajo la curva y sobre un solo punto es cero
c. $P(1.0 \le x \le 1.25) = 2(0.25) = 0.50$
d. $P(1.20 < x < 1.5) = 2(0.30) = 0.60$

2. b. 0.50
c. 0.60
d. 15
e. 8.33

4. a.



b. $P(.25 < x < 0.75) = 1(0.50) = 0.50$
c. $P(x \le 0.30) = 1(0.30) = 0.30$
d. $P(x > 0.60) = 1(0.40) = 0.40$

6. a. 0.40
b. 0.64
c. 0.68

10. a. 0.9332
b. 0.8413
c. 0.0919
d. 0.4938

12. a. 0.2967
b. 0.4418
c. 0.3300
d. 0.5910
e. 0.8849
f. 0.2389

13. a. $P(-1.98 \le z \le 0.49) = P(z \le 0.49) - P(z < -1.98) = 0.6879 - 0.0239 = 0.6640$
b. $P(0.52 \le z \le 1.22) = P(z \le 1.22) - P(z < 0.52) = 0.8888 - 0.6985 = 0.1903$
c. $P(-1.75 \le z \le -1.04) = P(z \le -1.04) - P(z < -1.75) = 0.1492 - 0.0401 = 0.1091$

14. a. $z = 1.96$
b. $z = 1.96$
c. $z = 0.61$
d. $z = 1.12$
e. $z = 0.44$
f. $z = 0.44$

15. a. El valor z que corresponde a la probabilidad acumulada de 0.2119 es $z = -0.80$

b. Se calcula $0.9030/2 = 0.4515$; la probabilidad acumulada $0.5000 + 0.4515 = 0.9414$ corresponde a $z = 1.66$

c. Se calcula $0.2052/2 = 0.1026$; la probabilidad acumulada $0.5000 + 0.1026 = 0.6026$ corresponde a $z = 0.26$

d. El valor z que corresponde a la probabilidad acumulada de 0.9948 es $z = 2.56$

e. El área a la izquierda de z es $1 - 0.6915 = 0.3085$, por lo que $z = 0.50$.

16. a. $z = 2.33$

b. $z = 1.96$

c. $z = 1.645$

d. $z = 1.28$

18. $\mu = 30$ y $\sigma = 8.2$

a. Para $x = 40$, $z = \dfrac{40 - 30}{8.2} = 1.22$

$P(z \leq 1.22) = 0.8888$

$P(x \geq 40) = 1.000 - 0.8888 = 0.1112$

b. Para $x = 20$, $z = \dfrac{20 - 30}{8.2} = -1.22$

$P(z \leq -1.22) = 0.1112$

$P(x \leq 20) = 0.1112$

c. El valor z 0.28 deja un área de aproximadamente 10% en la cola superior

$x = 30 + 8.2(1.28)$

$= 40.50$

Un precio por acción de por lo menos \$40.50 coloca a la empresa en el 10% de las mejores

20. a. 0.0885

b. 12.51%

c. 93.8 horas o más

22. a. 0.7193

b. \$35.59

c. 0.0233

24. a. 200, 26.04

b. 0.2206

c. 0.1251

d. 242.84 millones

26. a. $\mu = np = 100(0.20) = 20$

$\sigma^2 = np(1 - p) = 100(0.20)(0.80) = 16$

$\sigma = \sqrt{16} = 4$

b. Sí, porque $np = 20$ y $n(1 - p) = 80$

c. $P(23.5 \leq x \leq 24.5)$

$z = \dfrac{24.5 - 20}{4} = 1.13 \qquad P(z \leq 1.13) = 0.8708$

$z = \dfrac{23.5 - 20}{4} = 0.88 \qquad P(z \leq 0.88) = 0.8106$

$P(23.5 \leq x \leq 24.5) = P(0.88 \leq z \leq 1.13)$

$= 0.8708 - 0.8106 = 0.0602$

d. $P(17.5 \leq x \leq 22.5)$

$z = \dfrac{22.5 - 20}{4} = 0.63 \qquad P(z \leq 0.63) = 0.7357$

$z = \dfrac{17.5 - 20}{4} = -0.63 \qquad P(z \leq -0.63) = 0.2643$

$P(17.5 \leq x \leq 22.5) = P(-0.63 \leq z \leq 0.63)$

$= 0.7357 - 0.2643 = 0.4714$

e. $P(x \leq 15.5)$

$z = \dfrac{15.5 - 20}{4} = -1.13 \qquad P(z \leq -1.13) = 0.1292$

$P(x \leq 15.5) = P(z \leq -1.13) = 0.1292$

28. a. Al responder esta pregunta se supone que no se conoce la cantidad exacta de republicanos y demócratas

$\mu = np = 250(0.47) = 117.5$

$\sigma^2 = np(1 - p) = 250(0.47)(0.53) = 62.275$

$\sigma = \sqrt{62.275} = 7.89$

La mitad del grupo son 125 personas, de manea que se quiere hallar

$P(x \geq 124.5)$

Para $x = 124.5$, $z = \dfrac{124.5 - 117.5}{7.89} = 0.89$

$P(z \geq 0.89) = 1 - 0.8133 = 0.1867$

Por tanto, $P(x \geq 124.5) = 0.1867$

Se estima que la probabilidad de que por lo menos la mitad del grupo esté a favor de la propuesta es 0.1867

b. Para los republicanos: $np = 150(0.64) = 96$

Para los demócratas: $np = 100(0.29) = 29$

El número esperado a favor de la propuesta es $= 96 + 29 = 125$

c. De acuerdo con el inciso b se observa que se puede esperar exactamente la misma cantidad de personas a favor como en contra

30. a. 220

b. 0.0392

c. 0.8962

32. a. 0.5276

b. 0.3935

c. 0.4724

d. 0.1341

33. a. $P(x \leq x_0) = 1 - e^{-x_0/3}$

b. $P(x \leq 2) = 1 - e^{-2/3} = 1 - 0.5134 = 0.4866$

c. $P(x \geq 3) = 1 - P(x \leq 3) = 1 - (1 - e^{-3/3})$

$= e^{-1} = 0.3679$

d. $P(x \leq 5) = 1 - e^{-5/3} = 1 - 0.1889 = 0.8111$

e. $P(2 \leq x \leq 5) = P(x \leq 5) - P(x \leq 2)$

$= 0.8111 - 0.4866 = 0.3245$

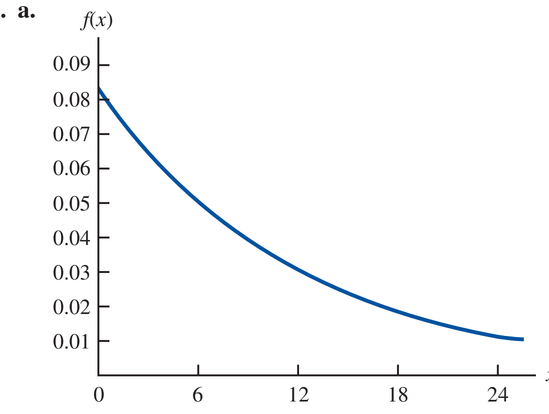
34. a. 0.5624

b. 0.1915

c. 0.2461

d. 0.2259

35. a.



b. $P(x \leq 12) = 1 - e^{-12/12} = 1 - 0.3679 = 0.6321$
c. $P(x \leq 6) = 1 - e^{-6/12} = 1 - 0.6065 = 0.3935$
d. $P(x \geq 30) = 1 - P(x < 30)$
$= 1 - (1 - e^{-30/12})$
$= 0.0821$

36. a. 50 horas
b. 0.3935
c. 0.1353

38. a. $f(x) = 5.5e^{-5.5x}$
b. 0.2528
c. 0.6002

40. a. \$3 780 o menos
b. 19.22%
c. \$8 167.50

42. a. 3 229
b. 0.2244
c. \$12 382 o más

44. a. 0.0228
b. \$50

46. a. 38.3%
b. 3.59% mayor, 96.41% menor
c. 38.21%

48. $\mu = 19.23$ onzas

50. a. Una pérdida de \$240
b. 0.1788
c. 0.3557
d. 0.0594

52. a. $\frac{1}{7}$ minuto
b. $7e^{-7x}$
c. 0.0009
d. 0.2466

54. a. 2 minutos
b. 0.2212
c. 0.3935
d. 0.0821

Capítulo 7

1. a. AB, AC, AD, AE, BC, BD, BE, CD, CE, DE
b. Como hay 10 muestras, la probabilidad que tiene cada una es $\frac{1}{10}$
c. E y C porque 8 y 0 no se pueden emplear; 5 corresponde a E; 7 no se puede emplear; el 5 se salta porque E ya se tomó en la muestra; 3 corresponde a C; el 2 ya no se necesita porque ya se tiene una muestra de tamaño 2

2. 22, 147, 229, 289

3. 459, 147, 385, 113, 340, 401, 215, 2, 33, 348

4. a. Bell South, LSI Logic, General Electric
b. 120

6. 2 782, 493, 825, 1807, 289

8. Maryland, Iowa, Florida State, Virginia, Pittsburgh, Oklahoma

10. a. finita; **b.** infinita; **c.** infinita; **d.** infinita; **e.** finita

11. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{54}{6} = 9$
b. $s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}}$
$\Sigma(x_i - \bar{x})^2 = (-4)^2 + (-1)^2 + 1^2 + (-2)^2 + 1^2 + 5^2$
$= 48$
$s = \sqrt{\frac{48}{6 - 1}} = 3.1$

12. a. 0.50
b. 0.3667

13. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{465}{5} = 93$
b.

	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	94	+1	1
	100	+7	49
	85	−8	64
	94	+1	1
	92	−1	1
Totales	465	0	116

$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{116}{4}} = 5.39$

14. a. 0.45
b. 0.15
c. 0.45

16. a. 0.10
b. 20
c. 0.72

18. a. 200
b. 5
c. Normal con $E(\bar{x}) = 200$ y $\sigma_{\bar{x}} = 5$
d. La distribución de probabilidad de $\bar{x}$

19. a. La distribución muestral es normal con
$E(\bar{x}) = \mu = 200$
$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 50/\sqrt{100} = 5$
Para ± 5, $195 \leq \bar{x} \leq 205$
Usando la tabla para la probabilidad normal estándar:
Para $\bar{x} = 205$, $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{5}{5} = 1$
$P(z \leq 1) = 0.8413$
Para $\bar{x} = 195$, $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{-5}{5} = -1$
$P(z < -1) = 0.1587$
$P(195 \leq \bar{x} \leq 205) = 0.8413 - 0.1587 = 0.6826$
b. Para ± 10, $190 \leq \bar{x} \leq 210$
Usando la tabla para la probabilidad normal estándar:
Para $\bar{x} = 210$, $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{10}{5} = 2$
$P(z \leq 2) = 0.9772$

Para $\bar{x} = 190$, $z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \dfrac{-10}{5} = -2$

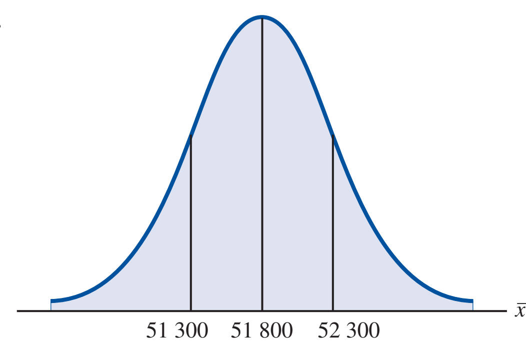
$P(z < -2) = 0.0228$

$P(190 \leq \bar{x} \leq 210) = 0.9722 - 0.0228 = 0.9544$

20. 3.54, 2.50, 2.04, 1.77
$\sigma_{\bar{x}}$ disminuye a medida que n aumenta

22. a. Normal con $E(\bar{x}) = 51{,}800$ y $\sigma_{\bar{x}} = 516.40$
b. $\sigma_{\bar{x}}$ disminuye a 365.15
c. $\sigma_{\bar{x}}$ disminuye a medida que n aumenta

23. a.



$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4\,000}{\sqrt{60}} = 516.40$$

Para $\bar{x} = 52\,300$, $z = \dfrac{52\,300 - 51\,800}{516.40} = 0.97$

$P(\bar{x} \leq 52\,300) = P(z \leq 0.97) = 0.8340$

Para $\bar{x} = 51\,300$, $z = \dfrac{51\,300 - 51\,800}{516.40} = -0.97$

$P(\bar{x} < 51\,300) = P(z < -0.97) = 0.1660$

$P(51\,300 \leq \bar{x} \leq 52\,300) = 0.8340 - 0.1660 = 0.6680$

b. $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{4\,000}{\sqrt{120}} = 365.15$

Para $\bar{x} = 52\,300$, $z = \dfrac{52\,300 - 51\,800}{365.15} = 1.37$

$P(\bar{x} \leq 52\,300) = P(z \leq 1.37) = 0.9147$

Para $\bar{x} = 51\,300$, $z = \dfrac{51\,300 - 51\,800}{365.15} = -1.37$

$P(\bar{x} < 51\,300) = P(z < -1.37) = 0.0853$

$P(51\,300 \leq \bar{x} \leq 52\,300) = 0.9147 - 0.0853 = 0.8294$

24. a. Normal con $E(\bar{x}) = 4\,260$ y $\sigma_{\bar{x}} = 127.28$
b. 0.95
c. 0.5704

26. a. 0.4246, 0.5284, 0.6922, 0.9586
b. Mayor probabilidad de que la media muestral esté cerca de la media poblacional

28. a. Normal con $E(\bar{x}) = 95$ y $\sigma_{\bar{x}} = 2.56$
b. 0.7580
c. 0.8502
d. Inciso c, tamaño de muestra mayor

30. a. $n/N = 0.01$; no
b. 1.29, 1.30; diferencia pequeña
c. 0.8764

32. a. $E(\bar{p}) = 0.40$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.40)(0.60)}{200}} = 0.0346$$

Como $\pm$ 0.03 significa $0.37 \leq \bar{p} \leq 0.43$

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.03}{0.0346} = 0.87$$

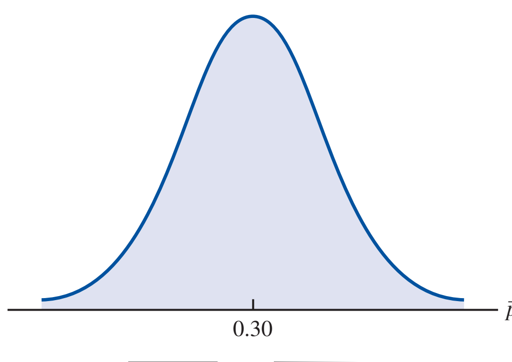
$P(0.37 \leq \bar{p} \leq 0.43) = P(-0.87 \leq z \leq 0.87)$
$= 0.8078 - 0.1922$
$= 0.6156$

b. $z = \dfrac{\bar{p} - p}{\sigma_{\bar{p}}} = \dfrac{0.05}{0.0346} = 1.44$

$P(0.35 \leq \bar{p} \leq 0.45) = P(-1.44 \leq z \leq 1.44)$
$= 0.9251 - 0.0749$

34. a. 0.6156
b. 0.7814
c. 0.9488
d. 0.9942
e. Mayor probabilidad a mayor n

35. a.



$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.30(0.70)}{100}} = 0.0458$$

La distribución normal es apropiada porque tanto $np = 100(0.30) = 30$ como $n(1 - p) = 100(0.70) = 70$ son mayores que 5

b. $P(0.20 \leq \bar{p} \leq 0.40) = ?$

$$z = \frac{0.40 - 0.30}{0.0458} = 2.18$$

$P(0.20 \leq \bar{p} \leq 0.40) = P(-2.18 \leq z \leq 2.18)$
$= 0.9854 - 0.0146$
$= 0.9708$

c. $P(0.25 \leq \bar{p} \leq 0.35) = ?$

$$z = \frac{0.35 - 0.30}{0.0458} = 1.09$$

$P(0.25 \leq \bar{p} \leq 0.35) = P(-1.09 \leq z \leq 1.09)$
$= 0.8621 - 0.1379$
$= 0.7242$

36. a. Normal con $E(\bar{p}) = 0.66$ y $\sigma_{\bar{p}} = 0.0273$
b. 0.8584
c. 0.9606
d. Sí, el error estándar es menor en el inciso c

e. 0.9616, la probabilidad es mayor porque el tamaño mayor de la muestra reduce el error estándar

38. a. Normal con $E(\bar{p}) = 0.56$ y $\sigma_{\bar{p}} = 0.0248$
b. 0.5820
c. 0.8926

40. a. Normal con $E(\bar{p}) = 0.76$ y $\sigma_{\bar{p}} = 0.0214$
b. 0.8384
c. 0.9452

42. 112, 145, 73, 324, 293, 875, 318, 618

44. a. Normal con $E(\bar{x}) = 115.50$ y $\sigma_{\bar{x}} = 5.53$
b. 0.9298
c. $z = -2.80$, 0.0026

46. a. 707
b. 0.50
c. 0.8414
d. 0.9544

50. a. Normal con $E(\bar{p}) = 0.28$ y $\sigma_{\bar{p}} = 0.0290$
b. 0.8324
c. 0.5098

52. a. 0.8882
b. 0.0233

54. a. 48
b. Normal con $E(\bar{p}) = 0.28$, $\sigma_{\bar{p}} = 0.0290$
c. 0.2119

Capítulo 8

2. Use $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$
a. $32 \pm 1.645(6/\sqrt{50})$
32 ± 1.4; 30.6 a 33.4
b. $32 \pm 1.96(6/\sqrt{50})$
32 ± 1.66; 30.34 a 33.66
c. $32 \pm 2.576(6/\sqrt{50})$
32 ± 2.19; 29.81 a 34.19

4. 54

5. a. $1.96\sigma/\sqrt{n} = 1.96(5/\sqrt{49}) = 1.40$
b. 24.80 ± 1.40; 23.40 a 26.20

6. 8.1 a 8.9

8. a. Que la población es por lo menos aproximadamente normal
b. 3.1
c. 4.1

10. a. \$113 638 a \$124 672
b. \$112 581 a \$125 729
c. \$110 515 a \$127 795
La amplitud aumenta a medida que el nivel de confianza aumenta

12. a. 2.179
b. -1.676
c. 2.457
d. -1.708 y 1.708
e. -2.014 y 2.014

13. a. $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{80}{8} = 10$
b. $s = \sqrt{\dfrac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\dfrac{84}{7}} = 3.464$
c. $t_{.025}\left(\dfrac{s}{\sqrt{n}}\right) = 2.365\left(\dfrac{3.46}{\sqrt{8}}\right) = 2.9$
d. $\bar{x} \pm t_{0.025}\left(\dfrac{s}{\sqrt{n}}\right)$
10 ± 2.9 (7.1 a 12.9)

14. a. 21.5 a 23.5
b. 21.3 a 23.7
c. 20.9 a 24.1
d. Un mayor margen de error y un intervalo más amplio

15. $\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$
Intervalo de confianza de 90%; $gl = 64$ y $t_{0.05} = 1.669$
$19.5 \pm 1.669\left(\dfrac{5.2}{\sqrt{65}}\right)$
19.5 ± 1.08 (18.42 a 20.58)
Intervalo de confianza de 95%; $gl = 64$ y $t_{0.05} = 1.998$
$19.5 \pm 1.998\left(\dfrac{5.2}{\sqrt{65}}\right)$
19.5 ± 1.29 (18.21 a 20.79)

16. a. 1.69
b. 47.31 a 50.69
c. Menos horas y costos más elevados para United

18. a. 3.8
b. 0.84
c. 2.96 a 4.64
d. Una n mayor para la repetición

20. $\bar{x} = 22$; 21.48 a 22.52

22. a. 3.35
b. 2.40 a 4.30

24. a. Valor planeado para $\sigma = \dfrac{\text{Rango}}{4} = \dfrac{36}{4} = 9$
b. $n = \dfrac{z^2_{0.025}\sigma^2}{E^2} = \dfrac{(1.96)^2(9)^2}{(3)^2} = 34.57$; use $n = 35$
c. $n = \dfrac{(1.96)^2(9)^2}{(2)^2} = 77.79$; use $n = 78$

25. a. Use $n = \dfrac{z^2_{\alpha/2}\sigma^2}{E^2}$
$n = \dfrac{(1.96)^2(6.84)^2}{(1.5)^2} = 79.88$; use $n = 80$
b. $n = \dfrac{(1.645)^2(6.84)^2}{(2)^2} = 31.65$; use $n = 32$

26. a. 18
b. 35
c. 97

28. a. 343
b. 487
c. 840
d. n aumenta; no a 99% de confianza

30. 81

31. a. $\bar{p} = \frac{100}{400} = 0.25$

b. $\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{0.25(0.75)}{400}} = 0.0217$

c. $\bar{p} \pm z_{0.025}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

$0.25 \pm 1.96(0.0217)$

0.25 ± 0.0424; 0.2076 a 0.2924

32. a. 0.6733 a 0.7267
b. 0.6682 a 0.7318

34. 1 068

35. a. $\bar{p} = \frac{281}{611} = 0.4599$ (46%)

b. $z_{0.05}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 1.645\sqrt{\frac{0.4599(1-0.4599)}{611}} = 0.0332$

c. $\bar{p} \pm 0.0332$
0.4599 ± 0.0332 (0.4267 a 0.4931)

36. a. 0.23
b. 0.1716 a 0.2884

38. a. 0.1790
b. 0.0738, 0.5682 a 0.7158
c. 354

39. a. $n = \frac{z_{.025}^2 p^*(1-p^*)}{E^2} = \frac{(1.96)^2(0.156)(1-0.156)}{(0.03)^2}$
$= 562$

b. $n = \frac{z_{0.005}^2 p^*(1-p^*)}{E^2} = \frac{(2.576)^2(0.156)(1-0.156)}{(0.03)^2}$
$= 970.77$; use 971

40. 0.0267 (0.8333 a 0.8867)

42. a. 0.0442
b. 601, 1 068, 2 401, 9 604

44. a. 4.00
b. \$29.77 a \$37.77

46. a. 998
b. \$24 479 a \$26 455
c. \$93.5 millones
d. Sí; \$21.4 (30%) superior a *El mundo perdido*

48. a. 14 minutos
b. 13.38 a 14.62
c. 32 por día
d. Reducción de personal

50. 37

52. 176

54. a. 0.5420
b. 0.0508
c. 0.4912 a 0.5928

56. a. 0.8273
b. 0.7957 a 0.8589

58. a. 1 267
b. 1 509

60. a. 0.3101
b. 0.2898 a 0.3304
c. 8219; no, este tamaño de muestra es innecesariamente grande

Capítulo 9

2. a. H_0: $\mu \leq 14$
H_a: $\mu > 14$
b. No hay evidencias de que el nuevo plan aumente las ventas
c. La hipótesis de investigación $\mu > 14$ tiene respaldo; el nuevo plan aumenta las ventas

4. a. H_0: $\mu \geq 220$
H_a: $\mu < 220$

5. a. Rechazar H_0: $\mu \leq 56.2$ siendo verdadera
b. Aceptar H_0: $\mu \leq 56.2$ siendo falsa

6. a. H_0: $\mu \leq 1$
H_a: $\mu > 1$
b. Afirmar que $\mu > 1$ cuando esto no es verdad
c. Afirmar que $\mu \leq 1$ cuando esto no es verdad

8. a. Afirmar que $m < 220$ cuando esto no es verdad
b. Afirmar que $m \geq 220$ cuando esto no es verdad

10. a. $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{26.4 - 25}{6/\sqrt{40}} = 1.48$
b. Usando la tabla de la distribución normal estándar con $z = 1.48$: valor-$p = 1.0000 - 0.9306 = 0.0694$
c. Valor-$p > 0.01$, no rechazar H_0
d. Rechazar H_0 si $z \geq 2.33$
$1.48 < 2.33$, no rechazar H_0

11. a. $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{14.15 - 15}{3/\sqrt{50}} = -2.00$
b. Valor-$p = 2(0.0228) = 0.0456$
c. Valor-$p \leq 0.05$, rechazar H_0
d. Rechazar H_0 si $z \leq -1.96$ o $z \geq 1.96$
$-2.00 \leq -1.96$, rechazar H_0

12. a. 0.1056; no rechazar H_0
b. 0.0062; rechazar H_0
c. ≈ 0; rechazar H_0
d. 0.7967; no rechazar H_0

14. a. 0.3844; no rechazar H_0
b. 0.0074; rechazar H_0
c. 0.0836; no rechazar H_0

15. a. H_0: $\mu \geq 1\,056$
H_a: $\mu < 1\,056$
b. $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{910 - 1\,056}{1\,600/\sqrt{400}} = -1.83$
valor-$p = 0.0336$
c. Valor-$p \leq 0.05$, rechazar H_0; el reembolso medio de los declarantes “de última hora” es menor que \$1 056

d. Rechazar H_0 si $z \leq -1.645$
$-1.83 \leq -1.645$; rechazar H_0

16. a. H_0: $\mu \leq 895$
H_a: $\mu > 895$
b. 0.1170
c. No rechazar H_0
d. Recolectar más datos

18. a. H_0: $\mu = 4.1$
H_a: $\mu \neq 4.1$
b. -2.21, 0.0272
c. Rechazar H_0

20. a. H_0: $\mu \geq 32.79$
H_a: $\mu < 32.79$
b. -2.73
c. 0.0032
d. Rechazar H_0

22. a. H_0: $\mu = 8$
H_a: $\mu \neq 8$
b. 0.1706
c. No rechazar H_0
d. 7.83 a 8.97; sí

24. a. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{17 - 18}{4.5/\sqrt{48}} = -1.54$
b. Grados de libertad $= n - 1 = 47$
El área en la cola inferior está entre 0.05 y 0.10
El valor-p (para dos colas) está entre 0.10 y 0.20
Valor-p exacto = 0.1303
c. Valor-$p > 0.05$; no rechazar H_0
d. Como $gl = 47$, $t_{0.025} = 2.012$
Rechazar H_0 si $t \leq -2.012$ o $t \geq 2.012$
$t = -1.54$; no rechazar H_0

26. a. Entre 0.02 y 0.05; el valor-p exacto = 0.0397; rechazar H_0
b. Entre 0.01 y 0.02; el valor-p exacto = 0.0125; rechazar H_0
c. Entre 0.10 y 0.20; el valor-p exacto = 0.1285; no rechazar H_0

27. a. H_0: $\mu \geq 238$
H_a: $\mu < 238$
b. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{231 - 238}{80/\sqrt{100}} = -0.88$
Grados de libertad $= n - 1 = 99$
El valor-p está entre 0.10 y 0.20
Valor-p exacto = 0.1905
c. Valor-$p > 0.05$; no rechazar H_0
No se puede concluir que la prestación media semanal en el estado de Virginia es menor que la media nacional
d. $gl = 99$, $t_{0.05} = -1.66$
Rechazar H_0 si $t \leq -1.66$
$-0.88 > -1.66$; no rechazar H_0

28. a. H_0: $\mu \leq 3\,530$
H_a: $\mu > 3\,530$
b. Entre 0.005 y 0.01
Valor-p exacto = 0.0072
c. Rechazar H_0

30. a. H_0: $\mu = 600$
H_a: $\mu \neq 600$
b. Entre 0.20 y 0.40
Valor-p exacto = 0.2491
c. No rechazar H_0
d. Un tamaño de muestra más grande

32. a. H_0: $\mu = 10\,192$
H_a: $\mu \neq 10\,192$
b. Entre 0.02 y 0.05
Valor-p exacto = 0.0304
c. Rechazar H_0

34. a. H_0: $\mu = 2$
H_a: $\mu \neq 2$
b. 2.2
c. 0.52
d. Entre 0.20 y 0.40
Valor-p exacto = 0.2535
e. No rechazar H_0

36. a. $z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \dfrac{0.68 - 0.75}{\sqrt{\dfrac{0.75(1-0.75)}{300}}} = -2.80$
Valor-p = 0.0026
Valor-$p \leq 0.05$; rechazar H_0
b. $z = \dfrac{0.72 - 0.75}{\sqrt{\dfrac{0.75(1-0.75)}{300}}} = -1.20$
Valor-p = 0.1151
Valor-$p > 0.05$; no rechazar H_0
c. $z = \dfrac{0.70 - 0.75}{\sqrt{\dfrac{0.75(1-0.75)}{300}}} = -2.00$
Valor-p = 0.0228
Valor-$p \leq 0.05$; rechazar H_0
d. $z = \dfrac{0.77 - 0.75}{\sqrt{\dfrac{0.75(1-0.75)}{300}}} = 0.80$
Valor-p = 0.7881
Valor-$p > 0.05$; no rechazar H_0

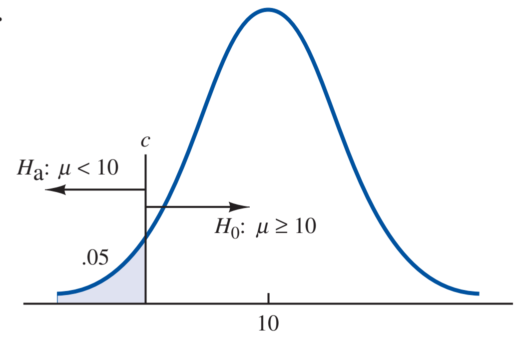
38. a. H_0: $p = 0.64$
H_a: $p \neq 0.64$
b. $\bar{p} = 52/100 = 0.52$
$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} = \dfrac{0.52 - 0.64}{\sqrt{\dfrac{0.64(1-0.64)}{100}}} = -2.50$
Valor-p = 2(0.0062) = 0.0124
c. Valor-$p \leq 0.05$; rechazar H_0
La proporción difiere del 0.64 reportado
d. Sí, ya que $\bar{p} = 0.52$ indica que pocos creen que la marca del supermercado sea tan buena como la otra marca

40. a. 0.2702
b. H_0: $p \leq 0.22$
H_a: $p > 0.22$
Valor-$p \approx 0$; rechazar H_0
c. Ayudan a evaluar la eficacia de los comerciales

42. H_0: $p \leq 0.24$
H_a: $p > 0.24$
Valor-p = 0.0023; rechazar H_0

44. **a.** H_0: $p \leq 0.51$
H_a: $p > 0.51$
b. $\bar{p} = 0.58$, valor-p = 0.0026
c. Rechazar H_0

46.



$$c = 10 - 1.645(5/\sqrt{120}) = 9.25$$
Rechazar H_0 si $\bar{x} < 9.25$

a. Si $\mu = 9$,
$$z = \frac{9.25 - 9}{5/\sqrt{120}} = 0.55$$
$$P(\text{Rechazar } H_0) = (1.0000 - 0.7088) = 0.2912$$
b. Error tipo II
c. Si $\mu = 8$,
$$z = \frac{9.25 - 8}{5/\sqrt{120}} = 2.74$$
$$\beta = (1.0000 - 0.9969) = 0.0031$$

48. **a.** Concluir $\mu \leq 15$ cuando no es verdad
b. 0.2676
c. 0.0179

49. **a.** H_0: $\mu \geq 25$
H_a: $\mu < 25$
Rechazar H_0 si $z < -2.05$
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 25}{3/\sqrt{30}} = -2.05$$
Hallar $\bar{x} = 23.88$
Regla de decisión: Aceptar H_0 si $\bar{x} > 23.88$
Rechazar H_0 si $\bar{x} \leq 23.88$
b. Si $\mu = 23$,
$$z = \frac{23.88 - 23}{3/\sqrt{30}} = 1.61$$
$$\beta = 1.0000 - 0.9463 = 0.0537$$
c. Si $\mu = 24$,
$$z = \frac{23.88 - 24}{3/\sqrt{30}} = -.22$$
$$\beta = 1.0000 - 0.4129 = 0.5871$$
d. En este caso no se puede cometer un error tipo II; observe que si $\mu = 25.5$, H_0 es verdadera; el error tipo II sólo se puede cometer cuando H_0 es falsa

50. **a.** Concluir que $\mu = 28$ cuando no es así
b. 0.0853, 0.6179, 0.6179, 0.0853
c. 0.9147

52. 0.1151, 0.0015
Al aumentar n se reduce β

54. $$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_0 - \mu_a)^2} = \frac{(1.645 + 1.28)^2(5)^2}{(10 - 9)^2} = 214$$

57. Para $\mu_0 = 400$, $\alpha = 0.02$; $z_{0.02} = 2.05$
Para $\mu_0 = 385$, $\beta = 0.10$; $z_{0.10} = 1.28$
Como $\sigma = 30$,
$$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_0 - \mu_a)^2} = \frac{(2.05 + 1.28)^2(30)^2}{(400 - 385)^2} = 44.4 \text{ o } 45$$

58. 324

60. **a.** H_0: $\mu = 16$
H_a: $\mu \neq 16$
b. 0.0286; rechazar H_0
reajustar línea
c. 0.2186; no rechazar H_0
Continuar operando
d. $z = 2.19$; rechazar H_0
$z = -1.23$; no rechazar H_0
Sí, la misma conclusión

62. **a.** H_0: $\mu \leq 119\,155$
H_a: $\mu > 119\,155$
b. 0.0047
c. Rechazar H_0

64. $t = -0.93$
Valor-p entre 0.20 y 0.40
Valor-p exacto = 0.3596
No rechazar H_0

66. $t = 2.26$
Valor-p entre 0.01 y 0.025
Valor-p exacto = 0.0155
Rechazar H_0

68. **a.** H_0: $p \leq 0.50$
H_a: $p > 0.50$
b. 0.64
c. 0.0026; rechazar H_0

70. **a.** H_0: $p \leq 0.80$
H_a: $p > 0.80$
b. 0.84
c. 0.0418
d. Rechazar H_0

72. H_0: $p \geq 0.90$
H_a: $p < 0.90$
Valor-p = 0.0808
No rechazar H_0

74. **a.** H_0: $\mu \leq 72$
H_a: $\mu > 72$
b. 0.2912
c. 0.7939
d. 0 porque H_0 es verdadera

76. a. 45
b. 0.0192, 0.2358, 0.7291, 0.7291, 0.2358, 0.0192

Capítulo 10

1. a. $\bar{x}_1 - \bar{x}_2 = 13.6 - 11.6 = 2$
b. $z_{\alpha/2} = z_{.05} = 1.645$

$$\bar{x}_1 - \bar{x}_2 \pm 1.645\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$2 \pm 1.645\sqrt{\frac{(2.2)^2}{50} + \frac{(3)^2}{35}}$$

2 ± 0.98 (1.02 a 2.98)
c. $z_{\alpha/2} = z_{.05} = 1.96$

$$2 \pm 1.96\sqrt{\frac{(2.2)^2}{50} + \frac{(3)^2}{35}}$$

2 ± 1.17 (0.83 a 3.17)

2. a. $$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(25.2 - 22.8) - 0}{\sqrt{\frac{(5.2)^2}{40} + \frac{(6)^2}{50}}} = 2.03$$
b. Valor-$p = 1.0000 - 0.9788 = 0.0212$
c. Valor-$p \leq 0.05$; rechazar H_0

4. a. $\bar{x}_1 - \bar{x}_2 = 2.04 - 1.72 = .32$
b. $$z_{0.025}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1.96\sqrt{\frac{(0.10)^2}{40} + \frac{(0.08)^2}{35}} = 0.04$$
c. 0.32 ± 0.04 (0.28 a 0.36)

6. Valor-$p = 0.015$
Rechazar H_0; un incremento

8. a. 1.08
b. 0.2802
c. No rechazar H_0; no se puede concluir que exista una diferencia

9. a. $\bar{x}_1 - \bar{x}_2 = 22.5 - 20.1 = 2.4$
b. $$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$= \frac{\left(\frac{2.5^2}{20} + \frac{4.8^2}{30}\right)^2}{\frac{1}{19}\left(\frac{2.5^2}{20}\right)^2 + \frac{1}{29}\left(\frac{4.8^2}{30}\right)^2} = 45.8$$

c. $gl = 45$, $t_{0.025} = 2.014$

$$t_{0.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.014\sqrt{\frac{2.5^2}{20} + \frac{4.8^2}{30}} = 2.1$$

d. 2.4 ± 2.1(0.3 a 4.5)

10. a. $$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(13.6 - 10.1) - 0}{\sqrt{\frac{5.2^2}{35} + \frac{8.5^2}{40}}} = 2.18$$
b. $$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$= \frac{\left(\frac{5.2^2}{35} + \frac{8.5^2}{40}\right)^2}{\frac{1}{34}\left(\frac{5.2^2}{35}\right)^2 + \frac{1}{39}\left(\frac{8.5^2}{40}\right)^2} = 65.7$$

Use $gl = 65$
c. $gl = 65$, el área en la cola se encuentra entre 0.01 y 0.025; valor-p para dos colas se encuentra entre 0.02 y 0.05 valor-p exacto ≤ 0.05; rechazar H_0

12. a. $\bar{x}_1 - \bar{x}_2 = 22.5 - 18.6 = 3.9$ millas
b. $$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$= \frac{\left(\frac{8.4^2}{50} + \frac{7.4^2}{40}\right)^2}{\frac{1}{49}\left(\frac{8.4^2}{50}\right)^2 + \frac{1}{39}\left(\frac{7.4^2}{40}\right)^2} = 87.1$$

Use $gl = 87$, $t_{0.025} = 1.988$

$$3.9 \pm 1.988\sqrt{\frac{8.4^2}{50} + \frac{7.4^2}{40}}$$

3.9 ± 3.3 (0.6 a 7.2)

14. a. H_0: $\mu_1 - \mu_2 = 0$
H_a: $\mu_1 - \mu_2 \neq 0$
b. 2.18
c. En la tabla t, el valor-p está entre 0.02 y 0.05
El valor-p exacto $= 0.03$
d. Rechazar H_0; las edades promedio son diferentes

16. a. H_0: $\mu_1 - \mu_2 \leq 0$
H_a: $\mu_1 - \mu_2 > 0$
b. 38
c. $t = 1.80$, $gl = 25$
En la tabla t, el valor-p está entre 0.025 y 0.05
El valor-p exacto = 0.0420
d. Rechazar H_0; se concluye que mejores puntuaciones si hay un nivel de enseñanza más alto

18. a. H_0: $\mu_1 - \mu_2 \geq 120$
H_a: $\mu_1 - \mu_2 < 120$
b. -2.10
En la tabla t, el valor-p está entre 0.01 y 0.025
El valor-p exacto = 0.0195
c. 32 a 118
d. Un tamaño de muestra más grande

19. a. 1, 2, 0, 0, 2
b. $\bar{d} = \Sigma d_i/n = 5/5 = 1$
c. $$s_d = \sqrt{\frac{\Sigma(d_i - \bar{d})^2}{n - 1}} = \sqrt{\frac{4}{5 - 1}} = 1$$
d. $$t = \frac{\bar{d} - \mu}{s_d/\sqrt{n}} = \frac{1 - 0}{1/\sqrt{5}} = 2.24$$

$gl = n - 1 = 4$
En la tabla t, el valor-p está entre 0.025 y 0.05
El valor-p exacto $= 0.0443$
Valor-$p \leq 0.05$; rechazar H_0

20. **a.** 3, −1, 3, 5, 3, 0, 1
b. 2
c. 2.08
d. 2
e. 0.07 a 3.93

21. H_0: $\mu_d \leq 0$
H_a: $\mu_d > 0$
$s_d = 1.30$
$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{.625 - 0}{1.30/\sqrt{8}} = 1.36$$
$gl = n - 1 = 7$
En la tabla t, el valor-p está entre 0.10 y 0.20
El valor-p exacto = 0.1080
Valor-p > 0.05; no rechazar H_0

22. \$0.10 a \$0.32

24. $t = 1.32$
En la tabla t, el valor-p es mayor que 0.10
El valor-p exacto = 0.1142
No rechazar H_0

26. **a.** $t = -0.60$
En la tabla t, el valor-p es mayor que 0.40
El valor-p exacto = 0.5633
No rechazar H_0
b. −0.103
c. 0.39; un tamaño de muestra mayor

28. **a.** $\bar{p}_1 - \bar{p}_2 = 0.48 - 0.36 = 0.12$
b. $\bar{p}_1 - \bar{p}_2 \pm z_{0.05}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$
$$0.12 \pm 1.645\sqrt{\frac{0.48(1-0.48)}{400} + \frac{0.36(1-0.36)}{300}}$$
0.12 ± 0.0614 (0.0586 a 0.1814)
c. $0.12 \pm 1.96\sqrt{\frac{0.48(1-0.48)}{400} + \frac{0.36(1-0.36)}{300}}$
0.12 ± 0.0731 (0.0469 a 0.1931)

29. **a.** $\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} = \frac{200(0.22)+300(0.16)}{200 + 300} = 0.1840$
$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
$$= \frac{0.22 - 0.16}{\sqrt{0.1840(1-0.1840)\left(\frac{1}{200} + \frac{1}{300}\right)}} = 1.70$$
Valor-p = 1.0000 − 0.9554 = 0.0446
b. Valor-$p \leq 0.05$; rechazar H_0

30. $\bar{p}_1 = 0.55$, $\bar{p}_2 = 0.48$
0.07 ± 0.0691

32. **a.** H_0: $p_w \leq p_m$
H_a: $p_w > p_m$
b. $\bar{p}_w = 0.3699$
c. $\bar{p}_m = 0.3400$
d. Valor-p = 0.1093
No rechazar H_0

34. **a.** 0.803
b. 0.849
c. H_0: $p_1 - p_2 \geq 0$
H_a: $p_1 - p_2 < 0$
d. Valor-p = 0.0055
Rechazar H_0

36. **a.** H_0: $p_1 - p_2 = 0$
H_a: $p_1 - p_2 \neq 0$
b. 0.13
c. Valor-p = 0.0404

38. **a.** H_0: $\mu_1 - \mu_2 = 0$
H_a: $\mu_1 - \mu_2 \neq 0$
$z = 2.79$
Valor-p = 0.0052
Rechazar H_0

40. **a.** H_0: $\mu_1 - \mu_2 \leq 0$
H_a: $\mu_1 - \mu_2 > 0$
b. $t = 0.60$, $gl = 57$
En la tabla t, el valor-p es mayor a 0.20
El valor-p exacto = 0.2754
No rechazar H_0

42. **a.** 15 (o \$15 000)
b. 9.81 a 20.19
c. 11.5%

44. **a.** *Valor-p* ≈ 0, rechazar H_0
b. 0.0468 a 0.1332

46. **a.** 163, 66
b. 0.0804 a 0.2198
c. Sí

Capítulo 11

2. $s^2 = 25$
a. Para 19 grados de libertad, $\chi^2_{0.05} = 30.144$ y $\chi^2_{.95} = 10.117$
$$\frac{19(25)}{30.144} \leq \sigma^2 \leq \frac{19(25)}{10.117}$$
$$15.76 \leq \sigma^2 \leq 46.95$$
b. Para 19 grados de libertad, $\chi^2_{0.025} = 32.852$ y $\chi^2_{.975} = 8.907$
$$\frac{19(25)}{32.852} \leq \sigma^2 \leq \frac{19(25)}{8.907}$$
$$14.46 \leq \sigma^2 \leq 53.33$$
c. $3.8 \leq \sigma \leq 7.3$

4. **a.** 0.22 a 0.71
b. 0.47 a 0.87

6. **a.** 0.2205, 47.95, 6.92
b. 5.27 a 10.11

8. a. 0.00845
b. 0.092
c. 0.0042 a 0.0244
0.065 a 0.156

9. H_0: $\sigma^2 \leq 0.0004$
H_a: $\sigma^2 > 0.0004$
$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(30-1)(0.0005)}{0.0004} = 36.25$$
Para 29 grados de libertad, el valor-p es mayor que 0.10
Valor-$p > 0.05$; no rechazar H_0
Las especificaciones del producto no parecen estarse violando

10. H_0: $\sigma^2 \leq 331.24$
H_a: $\sigma^2 > 331.24$
$\chi^2 = 52.07$, $gl = 35$
El valor-p está entre 0.025 y 0.05
Rechazar H_0

12. a. 0.8106
b. $\chi^2 = 9.49$
El valor-p es mayor a 0.20
No rechazar H_0

14. a. $F = 2.4$
El valor-p entre 0.025 y 0.05
Rechazar H_0
b. $F_{0.05} = 2.2$; rechazar H_0

15. a. La varianza muestral más grande es s_1^2
$$F = \frac{s_1^2}{s_2^2} = \frac{8.2}{4} = 2.05$$
Grados de libertad: 20, 25
En las tablas el área en la cola está entre 0.025 y 0.05
El valor-p para la prueba de dos colas está entre 0.05 y 0.10
El valor-$p > 0.05$; no rechazar H_0
b. Para una prueba de dos colas:
$F_{\alpha/2} = F_{0.025} = 2.30$
Rechazar H_0 si $F \geq 2.30$
$2.05 < 2.30$; no rechazar H_0

16. $F = 2.63$
El valor-p es menor a 0.01
Rechazar H_0

17. a. La población 1 es la de los automóviles de 4 años de antigüedad
H_0: $\sigma_1^2 \leq \sigma_2^2$
H_a: $\sigma_1^2 > \sigma_2^2$
b. $$F = \frac{s_1^2}{s_2^2} = \frac{170^2}{100^2} = 2.89$$
Grados de libertad: 25, 24
En las tablas el valor-p es menor que 0.01
El valor-$p \leq 0.01$; rechazar H_0
Se concluye que en los automóviles de 4 años de antigüedad la varianza en los costos anuales de reparación es mayor que en los de 2 años de antigüedad, lo que es de esperarse dado que es más probable que automóviles más viejos necesiten reparaciones más caras, lo que hace que la varianza en los costos anuales de reparación sean mayores

18. $F = 3.54$
El valor-p está entre 0.10 y 0.20
No rechazar H_0

20. $F = 5.29$
El valor-$p \approx 0$
Rechazar H_0

22. a. $F = 4$
El valor-p es menor que 0.01
Rechazar H_0

24. 10.72 a 24.68

26. a. $\chi^2 = 27.44$
El valor-p está entre 0. 01 y 0.025
Rechazar H_0
b. 0.00012 a 0.00042

28. $\chi^2 = 31.50$
El valor-p está entre 0. 05 y 0.10
Rechazar H_0

30. a. $n = 15$
b. 6.25 a 11.13

32. $F = 1.39$
No rechazar H_0

34. $F = 2.08$
El valor-p está entre 0. 05 y 0.10
Rechazar H_0

Capítulo 12

1. a. Frecuencias esperadas: $e_1 = 200(0.40) = 80$
$e_2 = 200(0.40) = 80$
$e_3 = 200(0.20) = 40$
Frecuencias observadas: $f_1 = 60$, $f_2 = 120$, $f_3 = 20$
$$\chi^2 = \frac{(60-80)^2}{80} + \frac{(120-80)^2}{80} + \frac{(20-40)^2}{40}$$
$$= \frac{400}{80} + \frac{1600}{80} + \frac{400}{40}$$
$$= 5 + 20 + 10 = 35$$
Grados de libertad: $k - 1 = 2$
$\chi^2 = 35$ indica que el valor-p es menor que 0.005
El valor-$p \leq 0.01$; rechazar H_0
b. Rechazar H_0 si $c^2 \geq 9.210$
$\chi^2 = 35$; rechazar H_0

2. $\chi^2 = 15.33$, $gl = 3$
El valor-p es menor que 0.005
Rechazar H_0

3. H_0: $p_{ABC} = 0.29$, $p_{CBS} = 0.28$, $p_{NBC} = 0.25$, $p_{IND} = 0.18$
H_a: Las proporciones no son
$p_{ABC} = 0.29$, $p_{CBS} = 0.28$, $p_{NBC} = 0.25$, $p_{IND} = 0.18$
Frecuencias esperadas: $300(0.29) = 87$, $300(0.28) = 84$
$300(0.25) = 75$, $300(0.18) = 54$
$e_1 = 87$, $e_2 = 84$, $e_3 = 75$, $e_4 = 54$

Frecuencias observadas: $f_1 = 95, f_2 = 70, f_3 = 89, f_4 = 46$

$$\chi^2 = \frac{(95-87)^2}{87} + \frac{(70-84)^2}{84} + \frac{(89-75)^2}{75} + \frac{(46-54)^2}{54} = 6.87$$

Grados de libertad: k − 1 = 3
$\chi^2 = 6.87$, el valor-p está entre 0.05 y 0.10
No rechazar H_0

4. $\chi^2 = 29.51$, $gl = 5$
El valor-p es menor que 0.005
Rechazar H_0

6. a. $\chi^2 = 12.21$, $gl = 3$
El valor-p está entre 0.005 y 0.01
Se concluye que hubo diferencia en el 2003
b. 21%, 30%, 15%, 34%
Aumento del uso de tarjeta de débito
c. 51%

8. $\chi^2 = 16.31$, $gl = 3$
El valor-p es menor que 0.005
Rechazar H_0

9. H_0: La variable de las columnas es independiente de la variable de los renglones
H_a: La variable de las columnas no es independiente de la variable de los renglones
Frecuencias esperadas:

	A	B	C
P	28.5	39.9	45.6
Q	21.5	30.1	34.4

$$\chi^2 = \frac{(20-28.5)^2}{28.5} + \frac{(44-39.9)^2}{39.9} + \frac{(50-45.6)^2}{45.6} + \frac{(30-21.5)^2}{21.5} + \frac{(26-30.1)^2}{30.1} + \frac{(30-34.4)^2}{34.4} = 7.86$$

Grados de libertad: (2 − 1)(3 − 1) = 2
$\chi^2 = 7.86$, el valor-p está entre 0.01 y 0.25
Rechazar H_0

10. $\chi^2 = 19.77$, $gl = 4$
El valor-p es menor que 0.005
Rechazar H_0

11. H_0: El tipo de boleto comprado es independiente del tipo de vuelo
H_a: El tipo de boleto comprado no es independiente del tipo de vuelo
Frecuencias esperadas:

$e_{11} = 35.59$ $e_{12} = 15.41$
$e_{21} = 150.73$ $e_{22} = 65.27$
$e_{31} = 455.68$ $e_{32} = 197.32$

Boleto	Vuelo	Frecuencia observada (f_i)	Frecuencia esperada (e_i)	$(f_i - e_i)^2/e_i$
Primera	Nacional	29	35.59	1.22
Primera	Internacional	22	15.41	2.82
Clase de negocios	Nacional	95	150.73	20.61
Clase de negocios	Internacional	121	65.27	47.59
Vuelo tradicional	Nacional	518	455.68	8.52
Vuelo tradicional	Internacional	135	197.32	19.68
Totales		920		$\chi^2 = 100.43$

Grados de libertad: (3 − 1)(2 − 1) = 2
$\chi^2 = 100.43$, el valor-p es menor que 0.005
Rechazar H_0

12. a. $\chi^2 = 7.95$; $gl = 3$
El valor-p está entre 0.025 y 0.05
Rechazar H_0
b. De 18 a 24 la usan más

14. a. $\chi^2 = 10.60$, $gl = 4$
El valor-p está entre 0.025 y 0.05
Rechazar H_0; no son independientes
b. El efecto negativo sobre las notas es mayor cuando las horas aumentan

16. a. $\chi^2 = 7.85$, $gl = 3$
El valor-p está entre 0.025 y 0.05
Rechazar H_0
b. Farmacéutica, 98.6%

18. $\chi^2 = 3.01$, $gl = 2$
El valor-p es mayor que 0.10
No rechazar H_0: 63.3%

20. Primero se estima μ a partir de los datos muestrales (tamaño de la muestra = 120)

$$\mu = \frac{0(39) + 1(30) + 2(30) + 3(18) + 4(3)}{120} = \frac{156}{120} = 1.3$$

Por tanto se usan las probabilidades de Poisson con $\mu = 1.3$ para calcular las frecuencias esperadas

x	Frecuencia observada	Probabilidad de Poisson	Frecuencia esperada	Diferencia ($f_i - e_i$)
0	39	0.2725	32.70	6.30
1	30	0.3543	42.51	−12.51
2	30	0.2303	27.63	2.37
3	18	0.0998	11.98	6.02
4 o más	3	0.0431	5.16	−2.17

$$\chi^2 = \frac{(6.30)^2}{32.70} + \frac{(-12.51)^2}{42.51} + \frac{(2.37)^2}{27.63} + \frac{(6.02)^2}{11.98} + \frac{(-2.17)^2}{5.16} = 9.04$$

Grados de libertad: 5 − 1 − 1 = 3
$\chi^2 = 9.04$, el valor-p está entre 0.025 y 0.05
Rechazar H_0; no es una distribución de Poisson

21. Como $n = 30$ se usarán seis clases con una probabilidad de 0.1667 para cada clase

$$\bar{x} = 22.8, \; s = 6.27$$

Los valores de z que crean seis intervalos, cada uno con una probabilidad de 0.1667 son –0.98, –0,43, 0, 0.43, 0.98

z	Valor de x
−0.98	22.8 − 0.98(6.27) = 16.66
−0.43	22.8 − 0.43(6.27) = 20.11
0	22.8 + 0.00(6.27) = 22.80
0.43	22.8 + 0.43(6.27) = 25.49
0.98	22.8 + 0.98(6.27) = 28.94

Intervalo	Frecuencia observada	Frecuencia esperada	Diferencia
menor que 16.66	3	5	−2
16.66–20.11	7	5	2
20.11–22.80	5	5	0
22.80–25.49	7	5	2
25.49–28.94	3	5	−2
28.94 y mayor	5	5	0

$$\chi^2 = \frac{(-2)^2}{5} + \frac{(2)^2}{5} + \frac{(0)^2}{5} + \frac{(2)^2}{5} + \frac{(-2)^2}{5} + \frac{(0)^2}{5}$$
$$= \frac{16}{5} = 3.20$$

Grados de libertad: 6 − 2 − 1 = 3
$\chi^2 = 3.20$, el valor-p es mayor que 0.10
No rechazar H_0
No se rechaza la suposición de una distribución normal

22. $\chi^2 = 4.302 \; gl = 2$
El valor-p es mayor que 0.10
No rechazar H_0

24. $\chi^2 = 2.8$, $gl = 3$
El valor-p es mayor que 0.10
No rechazar H_0

26. $\chi^2 = 8.04$, $gl = 3$
El valor-p está entre 0.025 y 0.05
Rechazar H_0

28. $\chi^2 = 4.64$, $gl = 2$
El valor-p está entre 0.05 y 0.10
No rechazar H_0

30. $\chi^2 = 42.53$, $gl = 4$
El valor-p es menor a 0.005
Rechazar H_0

32. $\chi^2 = 23.37$, $gl = 3$
El valor-p es menor que 0.005
Rechazar H_0

34. **a.** $\chi^2 = 12.86$, $gl = 2$
El valor-p es menor que 0.005
Rechazar H_0

b. 66.9, 30.3, 2.9
54.0, 42.0, 4.0

36. $\chi^2 = 6.17$, $gl = 6$
El valor-p es mayor que 0.10
No rechazar H_0

38. $\chi^2 = 7.75$, $gl = 3$
El valor-p está entre 0.05 y 0.10
No rechazar H_0

Capítulo 13

1. **a.** $\bar{\bar{x}} = (156 + 142 + 134)/3 = 144$

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$
$$= 6(156 - 144)^2 + 6(142 - 144)^2 + 6(134 - 144)^2$$
$$= 1\,488$$

b. $\text{CMTR} = \frac{\text{SCTR}}{k - 1} = \frac{1\,488}{2} = 744$

c. $s_1^2 = 164.4, \quad s_2^2 = 131.2, \quad s_3^2 = 110.4$

$$\text{SCE} = \sum_{j=1}^{k} (n_j - 1)s_j^2$$
$$= 5(164.4) + 5(131.2) + 5(110.4)$$
$$= 2030$$

d. $\text{CME} = \frac{\text{SCE}}{n_T - k} = \frac{2\,030}{18 - 3} = 135.3$

e.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Tratamientos	1488	2	744	5.50	0.0162
Error	2030	15	135.3		
Total	3518	17			

f. $F = \frac{\text{CMTR}}{\text{CME}} = \frac{744}{135.3} = 5.50$

De la tabla F (2 grados de libertad en el numerador y 15 grados de libertad en el denominador), el valor p está entre 0.01 y 0.025

Usando Excel o Minitab, el valor-p correspondiente a F = 5.50 es 0.0162

Como el valor-$p \leq \alpha = 0.05$, se rechaza la hipótesis de que las medias de los tres tratamientos son iguales

2.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Tratamientos	300	4	75	14.07	0.0000
Error	160	30	5.33		
Total	460	34			

4.

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	***F***	**Valor-*p***
Tratamientos	150	2	75	4.80	0.0233
Error	250	16	15.63		
Total	400	18			

Rechazar H_0 porque el valor-$p \leq \alpha = 0.05$

6. Como el valor-p = 0.0082 es menor que 0.05, se rechaza la hipótesis nula de que las medias de los tres tratamientos son iguales

8. $\bar{\bar{x}} = (79 + 74 + 66)/3 = 73$

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 6(79 - 73)^2 + 6(74 - 73)^2 + 6(66 - 73)^2 = 516$$

$$\text{CMTR} = \frac{\text{SSTR}}{k - 1} = \frac{516}{2} = 258$$

$$s_1^2 = 34 \qquad s_2^2 = 20 \qquad s_3^2 = 32$$

$$\text{SCE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 5(34) + 5(20) + 5(32) = 430$$

$$\text{CME} = \frac{\text{SCE}}{n_T - k} = \frac{430}{18 - 3} = 28.67$$

$$F = \frac{\text{CMTR}}{\text{CME}} = \frac{258}{28.67} = 9.00$$

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	***F***	**Valor-*p***
Tratamientos	516	2	258	9.00	0.003
Error	430	15	28.67		
Total	946	17			

De la tabla F (2 grados de libertad en el numerador y 15 grados de libertad en el denominador), el valor p es menor que 0.01

Usando Excel o Minitab, el valor-p correspondiente a F = 9.00 es 0.003

Como el valor-$p \leq \alpha = 0.05$, se rechaza la hipótesis de que las medias en las tres fábricas sean iguales; en otras palabras, el análisis de varianza lleva a la conclusión de que las medias de las puntuaciones obtenidas en los exámenes de las tres fábricas de NCP no son iguales.

10. El valor-p = 0.0000

Como el valor-$p \leq \alpha = 0.05$, se rechaza la hipótesis nula de que las medias de los tres grupos sean iguales

12. El valor-p = 0.0003

Como el valor-$p \leq \alpha = 0.05$, se rechaza la hipótesis nula de que las medias de millas por galón sean iguales en los tres automóviles

13. a. $\bar{\bar{x}} = (30 + 45 + 36)/3 = 37$

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 5(30 - 37)^2 + 5(45 - 37)^2 + 5(36 - 37)^2 = 570$$

$$\text{CMTR} = \frac{\text{SSTR}}{k - 1} = \frac{570}{2} = 285$$

$$\text{SCE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 4(6) + 4(4) + 4(6.5) = 66$$

$$\text{CME} = \frac{\text{SCE}}{n_T - k} = \frac{66}{15 - 3} = 5.5$$

$$F = \frac{\text{CMTR}}{\text{CME}} = \frac{285}{5.5} = 51.82$$

En la tabla F (2 grados de libertad en el numerador y 12 grados de libertad en el denominador), el valor-p es menor que 0.01

Usando Excel o Minitab, el valor-p correspondiente a F = 51.82 es 0.0000

Como el valor-$p \leq \alpha = 0.05$, se rechaza la hipótesis de que las medias en las tres poblaciones sean iguales

b. $\text{LSD} = t_{\alpha/2}\sqrt{\text{CME}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

$$= t_{0.025}\sqrt{5.5\left(\frac{1}{5} + \frac{1}{5}\right)}$$

$$= 2.179\sqrt{2.2} = 3.23$$

$|\bar{x}_1 - \bar{x}_2| = |30 - 45| = 15 > \text{LSD}$; diferencia significativa
$|\bar{x}_1 - \bar{x}_3| = |30 - 36| = 6 > \text{LSD}$; diferencia significativa
$|\bar{x}_2 - \bar{x}_3| = |45 - 36| = 9 > \text{LSD}$; diferencia significativa

c. $\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\text{CME}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

$$(30 - 45) \pm 2.179\sqrt{5.5\left(\frac{1}{5} + \frac{1}{5}\right)}$$

$$-15 \pm 3.23 = -18.23 \text{ a } -11.77$$

14. a. Significativa: valor-p = 0.0106

b. LSD = 15.34
1 y 2; significativa
1 y 3; no significativa
2 y 3; significativa

15. a.

	Fabricante 1	**Fabricante 2**	**Fabricante 3**
Media muestral	23	28	21
Varianza muestral	6.67	4.67	3.33

$\bar{\bar{x}} = (23 + 28 + 21)/3 = 24$

$$\text{SCTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$

$$= 4(23 - 24)^2 + 4(28 - 24)^2 + 4(21 - 24)^2$$

$$= 104$$

$$\text{CMTR} = \frac{\text{SCTR}}{k - 1} = \frac{104}{2} = 52$$

$$\text{SCE} = \sum_{j=1}^{k} (n_j - 1)s_j^2$$

$$= 3(6.67) + 3(4.67) + 3(3.33) = 44.0$$

$$\text{CME} = \frac{\text{CSE}}{n_T - k} = \frac{44.01}{12 - 3} = 4.89$$

$$F = \frac{\text{CMTR}}{\text{CME}} = \frac{52}{4.89} = 10.63$$

En la tabla F (2 grados de libertad en el numerador y 9 grados de libertad en el denominador), el valor-p es menor que 0.01

Usando Excel o Minitab, el valor-p correspondiente a $F = 10.63$ es 0.0043

Como el valor-$p \leq \alpha = 0.05$, se rechaza la hipótesis de que la media del tiempo necesario para mezclar un lote de un material sea la misma con las máquinas de los tres fabricantes

b. $\text{LSD} = t_{\alpha/2}\sqrt{\text{CME}\left(\frac{1}{n_1} + \frac{1}{n_3}\right)}$

$$= t_{0.025}\sqrt{4.89\left(\frac{1}{4} + \frac{1}{4}\right)}$$

$$= 2.262\sqrt{2.45} = 3.54$$

Como $|\bar{x}_1 - \bar{x}_3| = |23 - 21| = 2 < 3.54$, no parece haber una diferencia significativa entre las medias del fabricante 1 y el fabricante 3

16. $\bar{x}_1 - \bar{x}_2 \pm \text{LSD}$

$23 - 28 \pm 3.54$

$-5 \pm 3.54 = -8.54$ a -1.46

18. **a.** Significativa; valor-$p = 0.000$

b. Significativa; $2.3 > \text{LSD} = 1.19$

20. **a.** Significativa; valor-$p = 0.042$

b. LSD = 5.74; diferencia significativa entre embarcaciones pequeñas y grandes.

21. ***Medias de tratamiento***

$\bar{x}_{\cdot 1} = 13.6$, $\bar{x}_{\cdot 2} = 11.0$, $\bar{x}_{\cdot 3} = 10.6$

Medias de bloque

$\bar{x}_{1\cdot} = 9$, $\bar{x}_{2\cdot} = 7.67$, $\bar{x}_{3\cdot} = 15.67$, $\bar{x}_{4\cdot} = 18.67$, $\bar{x}_{5\cdot} = 7.67$

Otras medias

$\bar{\bar{x}} = 176/15 = 11.73$

Paso 1

$$\text{SCT} = \sum_i \sum_j (x_{ij} - \bar{\bar{x}})^2$$

$$= (10 - 11.73)^2 + (9 - 11.73)^2 + \cdots + (8 - 11.73)^2$$

$$= 354.93$$

Paso 2

$$\text{SCTR} = b \sum_j (\bar{x}_{\cdot j} - \bar{\bar{x}})^2$$

$$= 5[(13.6 - 11.73)^2 + (11.0 - 11.73)^2 + (10.6 - 11.73)^2] = 26.53$$

Paso 3

$$\text{SCBL} = k \sum_j (\bar{x}_{i\cdot} - \bar{\bar{x}})^2$$

$$= 3[(9 - 11.73)^2 + (7.67 - 11.73)^2 + (15.67 - 11.73)^2 + (18.67 - 11.73)^2 + (7.67 - 11.73)^2] = 312.32$$

Paso 4

SCE = SCT − SCTR − SCBL

= 354.93 − 26.53 − 312.32 = 16.08

Fuente de variación	Suma de cuadros	Grados de libertad	Cuadrado medio	*F*	Valor-*p*
Tratamientos	26.53	2	13.27	6.60	0.0203
Bloques	312.32	4	78.08		
Error	16.08	8	2.01		
Total	354.93	14			

De acuerdo con la tabla F (2 grados de libertad en el numerador y 8 en el denominador), el valor-p está entre 0.01 y 0.025

El valor-p exacto es 0.0203

Como el valor-$p \leq \alpha = 0.05$, se rechaza la hipótesis nula de que las medias de los tres tratamientos sean iguales

22.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	*F*	Valor-*p*
Tratamientos	310	4	77.5	17.69	0.0005
Bloques	85	2	42.5		
Error	35	8	4.38		
Total	430	14			

Significativa; valor-$p \leq \alpha = 0.05$

24. Valor-$p = 0.0453$

Como el valor-$p \leq \alpha = 0.05$, se rechaza la hipótesis nula de que tiempo en minutos que se necesita para afinar un motor sea el mismo con los dos analizadores

26. Significativa; valor-$p = 0.0000$

28. *Paso 1*

$$\text{SCT} = \sum_i \sum_j \sum_k (x_{ijk} - \bar{\bar{x}})^2$$

$$= (135 - 111)^2 + (165 - 111)^2 + \cdots + (136 - 111)^2 = 9\,028$$

Paso 2

$$\text{SCA} = br \sum_i (\bar{x}_{j\cdot} - \bar{\bar{x}})^2$$

$$= 3(2)[(104 - 111)^2 + (118 - 111)^2] = 588$$

Paso 3

$$\text{SCB} = ar \sum_j (\bar{x}_{\cdot j} - \bar{\bar{x}})^2$$

$$= 2(2)[(130 - 111)^2 + (97 - 111)^2 + (106 - 111)^2] = 2\,328$$

Paso 4

$$\text{SCAB} = r \sum_i \sum_j (\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2$$

$$= 2[(150 - 104 - 130 + 111)^2 + (78 - 104 - 97 + 111)^2 + \cdots + (128 - 118 - 106 + 111)^2] = 4\,392$$

Paso 5

SCE = SCT − SCA − SCB − SCAB

= 9 028 − 588 − 2 328 − 4 392 = 1 720

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Factor A	588	1	588	2.05	0.2022
Factor B	2328	2	1164	4.06	0.0767
Interacción	4392	2	2196	7.66	0.0223
Error	1720	6	286.67		
Total	9028	11			

Factor A: $F = 2.05$
De acuerdo con la tabla F (1 grado de libertad en el numerador y 6 grados de libertad en el denominador), el valor-p es mayor que 0.10
Usando Excel o Minitab, el valor-p correspondiente a $F = 2.05$ es 0.2022
Como el valor-$p > \alpha = 0.05$, el factor A no es significativo
Factor B: $F = 4.06$
De acuerdo con la tabla F (2 grados de libertad en el numerador y 6 grados de libertad en el denominador), el valor-p está entre 0.05 y 0.10
Usando Excel o Minitab, el valor-p correspondiente a $F = 4.06$ es 0.0767
Como el valor-$p > \alpha = 0.05$, el factor B no es significativo
Interacción: $F = 7.66$
De acuerdo con la tabla F (2 grados de libertad en el numerador y 6 grados de libertad en el denominador), el valor-p está entre 0.01 y 0.025
Usando Excel o Minitab, el valor-p correspondiente a $F = 7.66$ es 0.0223
Como el valor-$p \leq \alpha = 0.05$, la interacción es significativa

30. Diseño: valor-$p = 0.0104$; significativa
Tamaño: valor-$p = 0.1340$; no significativa
Interacción: valor-$p = 0.2519$; no significativa

32. Género: valor-$p = 0.0001$; significativa
Ocupación: valor-$p = 0.0001$; significativa
Interacción: valor-$p = 0.0106$; significativa

34. Significativa; valor-$p = 0.0134$

36. Significativa; valor-$p = 0.046$

38. No significativa; valor-$p = 0.2455$

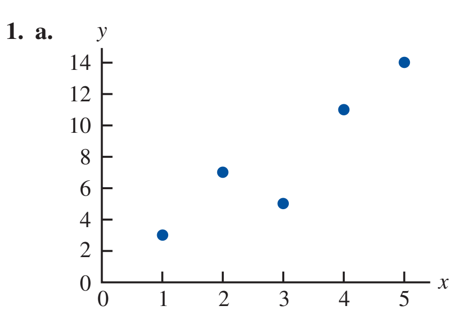
40. **a.** Significativa; valor-$p = 0.0175$

42. Significativa; valor-$p = 0.004$

44. Tipo de máquina (valor-$p = 0.0226$) es significativa; tipo de suministro (valor-$p = 0.7913$) e interacción (valor-$p = 0.0671$) no son significativas.

Capítulo 14

1. **a.**



b. Parece que existe una relación lineal positiva entre x y y

c. Se pueden trazar muchas líneas rectas para tratar de dar una aproximación lineal para la relación entre x y y; en el inciso d se determinará la ecuación de la recta que representa "mejor" la relación de acuerdo con el criterio de mínimos cuadrados

d. Sumatorias necesarias para calcular la pendiente y la intersección con el eje y

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{15}{5} = 3, \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{40}{5} = 8,$$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 26, \quad \Sigma(x_i - \bar{x})^2 = 10$$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{26}{10} = 2.6$$

$$b_0 = \bar{y} - b_1\bar{x} = 8 - (2.6)(3) = 0.2$$

$$\hat{y} = 0.2 - 2.6x$$

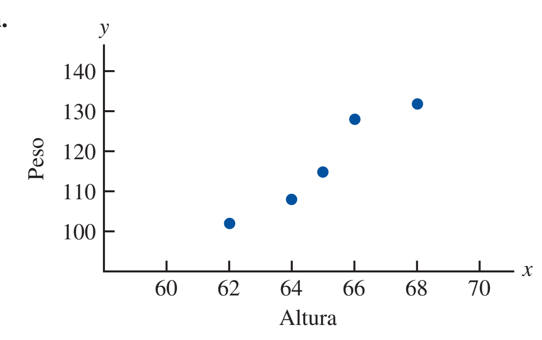
e. $\hat{y} = .2 + 2.6x = .2 + 2.6(4) = 10.6$

2. **b.** Parece que existe una relación lineal negativa entre x y y

d. $\hat{y} = 68 - 3x$

e. 38

4. **a.**



b. Para dar una relación lineal que aproxime la relación entre altura y peso

c. Se pueden trazar muchas líneas rectas para tratar de dar una aproximación lineal para la relación entre altura y peso; en el inciso d se determinará la ecuación de la recta que representa "mejor" esta relación de acuerdo con el criterio de mínimos cuadrados

d. Sumatorias necesarias para calcular la pendiente y la intersección con el eje y

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{325}{5} = 65, \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{585}{5} = 117,$$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 110, \quad \Sigma(x_i - \bar{x})^2 = 20$$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{110}{20} = 5.5$$

$$b_0 = \bar{y} - b_1\bar{x} = 117 - (5.5)(65) = -240.5$$

$$\hat{y} = -240.5 + 5.5x$$

e. $\hat{y} = -240.5 + 5.5(63) = 106$

6. **c.** $\hat{y} = -10.1641 + 0.1843x$

e. 11.95 o aproximadamente $12 000

8. **c.** $\hat{y} = 490.21 + 204.24x$

d. 1 307

10. c. $\hat{y} = 359.2668 - 5.2772x$
d. \$254

12. c. $\hat{y} = -8129.4439 + 22.4443x$
d. \$8 704

14. b. $\hat{y} = 28.30 - 0.0415x$
c. 26.2

15. a. $\hat{y}_i = 0.2 + 2.6x_i$ y $\bar{y} = 8$

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	3	2.8	0.2	0.04	−5	25
2	7	5.4	1.6	2.56	−1	1
3	5	8.0	−3.0	9.00	−3	9
4	11	10.6	0.4	0.16	3	9
5	14	13.2	0.8	0.64	6	36
				SCE = 12.40		SCT = 80

SCR = SCT − SCE = 80 − 12.4 = 67.6

b. $r^2 = \dfrac{\text{SCR}}{\text{SCT}} = \dfrac{67.6}{80} = 0.845$

La recta de mínimos cuadrados proporciona un buen ajuste; 84.5% de la variabilidad en y ha sido explicada por la recta de mínimos cuadrados

16. a. SCE = 230, SCT = 1850, SCR = 1 620
b. $r^2 = 0.876$
c. $r_{xy} = -0.936$

18. a. Ecuación de regresión estimada y media de la variable dependiente:

$\hat{y} = 1790.5 + 581.1x, \quad \bar{y} = 3\,650$

Suma de cuadrados debidos al error y suma total de cuadrados:

$\text{SCE} = \Sigma(y_i - \hat{y}_i)^2 = 85\,135.14$

$\text{SCT} = \Sigma(y_i - \bar{y})^2 = 335\,000$

Por tanto, SCR = SCT − SCE
= 335 000 − 85 135.14 = 249 864.86

b. $r^2 = \dfrac{\text{SCR}}{\text{SCRT}} = \dfrac{249\,864.86}{335\,000} = 0.746$

La recta de mínimos cuadrados explica 74.6% del total de la suma de cuadrados

20. a. $\hat{y} = 12.0169 + 0.0127x$
b. $r^2 = 0.4503$
c. 53

22. a. $\hat{y} = -745.480627 + 117.917320x$
b. $r^2 = 0.7071$
c. $r_{xy} = +0.84$

23. a. $s^2 = \text{CME} = \dfrac{\text{SCE}}{n-2} = \dfrac{12.4}{3} = 4.133$

b. $s = \sqrt{\text{CME}} = \sqrt{4.133} = 2.033$

c. $\Sigma(x_i - \bar{x})^2 = 10$

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{2.033}{\sqrt{10}} = 0.643$$

d. $t = \dfrac{b_1 - \beta_1}{s_{b_1}} = \dfrac{2.6 - 0}{0.643} = 4.044$

De acuerdo con la tabla t (3 grados de libertad) el área en la cola está entre 0.01 y 0.25

El valor-p está entre 0.02 y 0.05

Usando Excel o Minitab, el valor-p correspondiente a $t = 4.04$ es 0.0272

Como el valor-$p \leq \alpha$, se rechaza H_0: $\beta_1 = 0$

e. $\text{CMR} = \dfrac{\text{SCR}}{1} = 67.6$

$$F = \frac{\text{CMR}}{\text{CME}} = \frac{67.6}{4.133} = 16.36$$

De acuerdo con la tabla F (1 grado de libertad en el numerador y 3 en el denominador) el valor-p está entre 0.025 y 0.05

Usando Excel o Minitab, el valor-p correspondiente a F = 16.36 es 0.0272

Como el valor-$p \leq \alpha$, se rechaza H_0: $\beta_1 = 0$

Fuente de Variación	Sumas de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Regresión	67.6	1	67.6	16.36	0.0272
Error	12.4	3	4.133		
Totales	80	4			

24. a. 76.6667
b. 8.7560
c. 0.6526
d. Significativa; valor-p = 0.0193
e. Significativa; valor-p = 0.0193

26. a. $s^2 = \text{CME} = \dfrac{\text{SCE}}{n-2} = \dfrac{85\,135.14}{4} = 21\,283.79$

$s = \sqrt{\text{CME}} = \sqrt{21\,283.79} = 145.89$

$\Sigma(x_i - \bar{x})^2 = 0.74$

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{145.89}{\sqrt{.74}} = 169.59$$

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{581.08 - 0}{169.59} = 3.43$$

De acuerdo con la tabla t (4 grados de libertad) el área en la cola está entre 0.01 y 0.25

El valor-p está entre 0.02 y 0.05

Usando Excel o Minitab, el valor-p correspondiente a t = 3.43 es 0.0266

Como el valor-$p \leq \alpha$, se rechaza H_0: $\beta_1 = 0$

b. $\text{CMR} = \dfrac{\text{SCR}}{1} = \dfrac{249\,864.86}{1} = 249\,864.86$

$$F = \frac{\text{CMR}}{\text{CME}} = \frac{249\,864.86}{21\,283.79} = 11.74$$

De acuerdo con la tabla F (1 grado de libertad en el numerador y 4 en el denominador) el valor-p está entre 0.025 y 0.05

Usando Excel o Minitab, el valor-p correspondiente a F = 11.74 es 0.0266

Como el valor-$p \leq \alpha$, se rechaza H_0: $\beta_1 = 0$

c.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	Valor-p
Regresión	29 864.86	1	29 864.86	11.74	0.0266
Error	85 135.14	4	21 283.79		
Total	335 000	5			

28. Están relacionados: valor-p = 0.000

30. Significativa; valor-p = 0.002

32. **a.** $s = 2.033$

$\bar{x} = 3,\ \Sigma(x_i - \bar{x})^2 = 10$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

$$= 2.033\sqrt{\frac{1}{5} + \frac{(4-3)^2}{10}} = 1.11$$

b. $\hat{y} = 0.2 + 2.6x = 0.2 + 2.6(4) = 10.6$

$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$

$10.6 \pm 3.182(1.11)$

10.6 ± 3.53 o 7.07 a 14.13

c. $$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

$$= 2.033\sqrt{1 + \frac{1}{5} + \frac{(4-3)^2}{10}} = 2.32$$

d. $\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}}$

$10.6 \pm 3.182(2.32)$

10.6 ± 7.38 o 3.22 a 17.98

34. Intervalo de confianza: 8.65 a 21.15

Intervalo de predicción: −4.50 a 41.30

35. **a.** $s = 145.89,\ \bar{x} = 3.2,\ \Sigma(x_i - \bar{x})^2 = 0.74$

$\hat{y} = 1\,790.5 + 581.1x = 1\,790.5 + 581.1(3)$

$= 3\,533.8$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

$$= 145.89\sqrt{\frac{1}{6} + \frac{(3-3.2)^2}{0.74}} = 68.54$$

$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$

$3\,533.8 \pm 2.776(68.54)$

$3\,533.8 \pm 190.27$, o \$3 343.53 a \$3 724.07

b. $$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

$$= 145.89\sqrt{1 + \frac{1}{6} + \frac{(3-3.2)^2}{0.74}} = 161.19$$

$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}}$

$3\,533.8 \pm 2.776(161.19)$

$3\,533.8 \pm 447.46$, o \$3 086.34 a \$3 981.26

36. **a.** \$201

b. 167.25 a 234.65

c. 108.75 a 293.15

38. **a.** \$5 046.67

b. \$3 815.10 a \$6 278.24

c. No son fuera de lo común

40. **a.** 9

b. $\hat{y} = 20.0 + 7.21x$

c. 1.3626

d. SCE = SCT − SCR = 51 984.1 − 41 587.3 = 10 396.8

CME = 10 396.8/7 = 1485.3

$$F = \frac{\text{CMR}}{\text{CME}} = \frac{41{,}587.3}{1485.3} = 28.0$$

De acuerdo con la tabla F (1 grado de libertad en el numerador y 7 en el denominador) el valor-p es menor que 0.01

Usando Excel o Minitab, el valor-p correspondiente a F = 28.0 es 0.0011 es 0.02966

Como el valor-$p \leq \alpha = 0.05$, se rechaza H_o: $\beta_1 = 0$

e. $\hat{y} = 20.0 + 7.21(50) = 380.5$ o \$380 500

42. **a.** $\hat{y} = 80.0 + 50.0x$

b. 30

c. Significativa; valor-p = 0.000

d. \$680 000

44. **b.** Sí

c. $\hat{y} = 37.1 - 0\ \ 0.779x$

d. Significativa; valor-p = 0.003

e. $r^2 = 0.434$; no proporciona un buen ajuste

f. \$12.27 a \$22.90

g. \$17.47 a \$39.05

45. **a.** $$\bar{x} = \frac{\Sigma x_i}{n} = \frac{70}{5} = 14,\quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{76}{5} = 15.2,$$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 200,\ \Sigma(x_i - \bar{x})^2 = 126$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{200}{126} = 1.5873$$

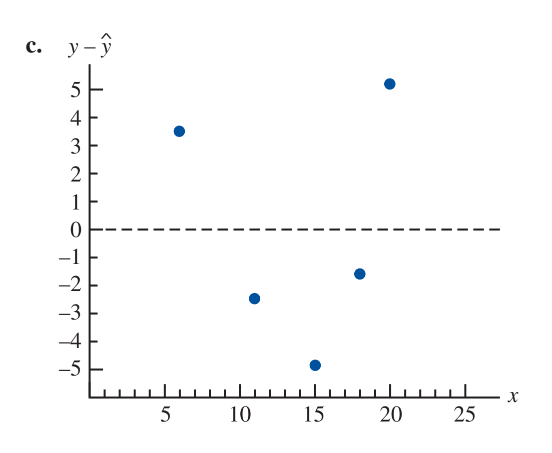
$b_0 = \bar{y} - b_1\bar{x} = 15.2 - (1.5873)(14) = -7.0222$

$\hat{y} = -7.02 + 1.59x$

b.

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
6	6	2.52	3.48
11	8	10.47	−2.47
15	12	16.83	−4.83
18	20	21.60	−1.60
20	30	24.78	5.22

c.



Con sólo cinco observaciones, es difícil determinar si se satisfacen las suposiciones; sin embargo, la gráfica sugiere curvatura en los residuales, lo que indicaría que las suposiciones sobre el término del error no se satisfacen; el diagrama de dispersión de estos datos indica también que la relación entre x y y puede que sea curvilínea

d. $s^2 = 23.78$

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$$

$$= \frac{1}{5} + \frac{(x_i - 14)^2}{126}$$

x_i	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Residuales estandarizados
6	0.7079	2.64	3.48	1.32
11	0.2714	4.16	−2.47	−0.59
15	0.2079	4.34	−4.83	−1.11
18	0.3270	4.00	−1.60	−0.40
20	0.4857	3.50	5.22	1.49

e. La gráfica de los residuales estandarizados contra $\hat{y}$ tiene la misma forma que la gráfica original de residuales; como se dijo en el inciso c, la curvatura observada indica que las suposiciones respecto al término del error puede que no se satisfagan

46. a. $\hat{y} = 2.32 + 0.64x$

b. No; la varianza no parece ser la misma para todos los valores de x

47. a. Sea x = gastos en publicidad y y = ingresos

b. SCT = 1002, SCE = 310.28, SCR = 691.72

$$\text{CMR} = \frac{\text{SCR}}{1} = 691.72$$

$$\text{CME} = \frac{\text{SCE}}{n-2} = \frac{310.28}{5} = 62.0554$$

$$F = \frac{\text{CMR}}{\text{CME}} = \frac{691.72}{62.0554} = 11.15$$

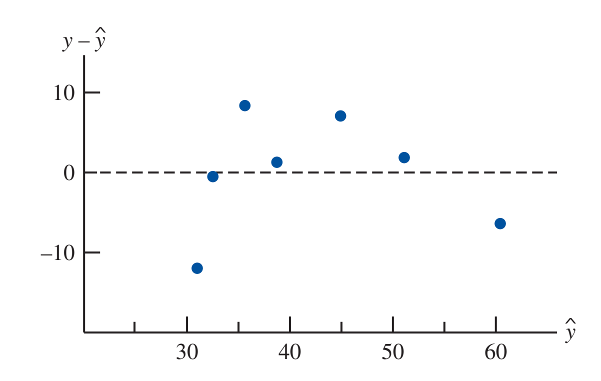
De acuerdo con la tabla F (1 grado de libertad en el numerador y 5 en el denominador) el valor-p está entre 0.01 y 0.025

Usando Excel o Minitab, el valor-p = 0.0206

Como el valor-$p \leq \alpha = 0.05$, se concluye que las dos variables están relacionadas

c.

x_i	y_i	$\hat{y}_i = 29.40 + 1.55x_i$	$y_i - \hat{y}_i$
1	19	30.95	−11.95
2	32	32.50	−0.50
4	44	35.60	8.40
6	40	38.70	1.30
10	52	44.90	7.10
14	53	51.10	1.90
20	54	60.40	−6.40



d. La gráfica de residuales lleva a cuestionarse la suposición de una relación lineal entre x y y; aun cuando la relación sea significativa al nivel $\alpha = 0.05$, sería extremadamente peligroso extrapolar más allá del intervalo en el que se encuentran los datos

48. b. Sí

50. a. Usando Minitab se obtiene la ecuación de regresión estimada $\hat{y} = 66.1 + 0.402x$; en la figura D14.50 se muestra parte de los resultados que da Minitab. Se presentan los valores ajustados y los residuales estandarizados:

x_i	y_i	$\hat{y}_i$	Residuales estandarizados
135	145	120.41	2.11
110	100	110.35	−1.08
130	120	118.40	0.14
145	120	124.43	−0.38
175	130	136.50	−0.78
160	130	130.47	−0.04
120	110	114.38	−0.41

b.

Residuales estandarizados

2.5 2.0 1.5 1.0 0.5 0.0 −0.5 −1.0 −1.5

105 110 115 120 125 130 135 140 $\hat{y}$

La gráfica de residuales estandarizados indica que la observación $x = 135$, $y = 145$ puede ser una observación atípica; note que esta observación tiene un residual estandarizado de 2.11

FIGURA D14.50

```
The regression equation is
Y = 66.1 + 0.402 X

Predictor      Coef    SE Coef       T       p
Constant      66.10      32.06    2.06   0.094
X            0.4023     0.2276    1.77   0.137

S = 12.62     R-sq = 38.5%     R-sq(adj) = 26.1%

Analysis of Variance

SOURCE            DF        SS       MS       F        p
Regression         1     497.2    497.2    3.12    0.137
Residual Error     5     795.7    159.1
Total              6    1292.9

Unusual Observations
Obs      X         Y       Fit    SE Fit    Residual    St Resid
  1    135    145.00    120.42      4.87       24.58       2.11R

 R denotes an observation with a large standardized residual
```

FIGURA D14.52

```
The regression equation is
Shipment = 4.09 + 0.196 Media$

Predictor       Coef    SE Coef       T       p
Constant       4.089      2.168    1.89   0.096
Media$       0.19552    0.03635    5.38   0.000

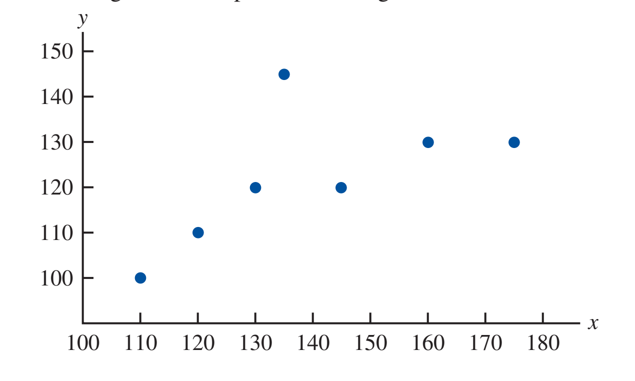
S = 5.044     R-Sq = 78.3%     R-Sq(adj) = 75.6%

Analysis of Variance
Source            DF       SS        MS        F        p
Regression         1   735.84    735.84    28.93    0.000
Residual Error     8   203.51     25.44
Total              9   939.35

Unusual Observations
Obs    Media$    Shipment      Fit    SE Fit    Residual    St Resid
  1       120       36.30    27.55      3.30        8.75       2.30R

R denotes an observation with a large standardized residual
```

c. El diagrama de dispersión es el siguiente:



El diagrama de dispersión también indica que la observación $x = 135$, $y = 145$ es una observación atípica; en la regresión lineal simple, las observaciones atípicas pueden identificarse observando el diagrama de dispersión

52. a. En la figura D14.52 se muestra una parte del resultado que da Minitab

b. Minitab identifica la observación 1 como una observación que tiene un residual estandarizado grande; por tanto se considerará a la observación 1 como una observación atípica

54. a. $\hat{y} = 707 + .00482x$

b. La observación 6 es una observación influyente

58. **a.** $\hat{y} = 9.26 + 0.711x$
b. Significativa; valor-p = 0.001
c. $r^2 = 0.744$; buen ajuste
d. \$13.53

60. **b.** $\hat{y} = -182.11 + 133428$ DJIA
c. Significativa; *valor*-p = 0.000
d. Excelente ajuste; $r^2 = 0.956$
e. 1 286

62. **a.** $\hat{y} = 22.2 - 0.148x$
b. Relación significativa; *valor*-p = 0.028
c. Buen ajuste; $r^2 = 0.739$
d. 12.294 a 17.271

64. **a.** $\hat{y} = 220 + 132x$
b. Significativa; *valor*-p = 0.000
c. $r^2 = 0.873$; muy buen ajuste
d. \$559.50 a \$933.90

66. **a.** Beta del mercado = 0.95
b. Significativa; *valor*-p = 0.029
c. $r^2 = 0.470$; sin buen ajuste
d. Texas Instrument tiene mayor riesgo

68. **b.** Parece que existe una relación lineal positiva entre las dos variables
c. $\hat{y} = 9.37 + 1.2875$ Cinco mejores (%)
d. Significativa; valor-p = 0.000
e. $r^2 = 0.741$; buen ajuste
f. $r_{xy} = 0.86$

Capítulo 15

2. **a.** La ecuación de regresión estimada es
$\hat{y} = 45.06 + 1.94x_1$
La estimación de y cuando $x_1 = 45$ es
$\hat{y} = 45.06 + 1.94(45) = 132.36$
b. La ecuación de regresión estimada es
$\hat{y} = 85.22 + 4.32x_2$
La estimación de y cuando $x_2 = 15$ es
$\hat{y} = 85.22 + 4.32(15) = 150.02$
c. La ecuación de regresión estimada es
$\hat{y} = -18.37 + 2.01x_1 + 4.74x_2$
La estimación de y cuando $x_1 = 45$ y $x_2 = 15$ es
$\hat{y} = -18.37 + 2.01(45) + 4.74(15) = 143.18$

4. **a.** \$255 000

5. **a.** El resultado de Minitab se muestra en la figura D15.5a
b. El resultado de Minitab se muestra en la figura D15.5b
c. En el inciso a es 1.60 y en el inciso b es 2.29; en el inciso a el coeficiente es una estimación de la variación en el ingreso debida a una variación de una unidad en los gastos en publicidad en televisión; en el inciso b el coeficiente representa una estimación de la variación en el ingreso debida a una variación de una unidad en los gastos en publicidad en televisión cuando la cantidad de publicidad en periódicos permanece constante
d. Ingreso = 83.2 + 2.29(3.5) + 1.30(1.8) = 93.56 o \$93 560

6. **a.** Proporción de ganados = 0.354 + 0.000888 HR
b. Proporción de ganados = 0.865 − 0.0837 ERA
c. Proporción de ganados = 0.709 + 0.00140 HR − 0.103 ERA

8. **a.** Ingreso = 247 – 32.8 Seguridad + 34.6 Coeficiente de gastos
b. 70.2

10. **a.** PCT = −1.22 + 3.96 FG%
b. aumento de 1% en FG% incrementará PCT en 0.04
c. PCT = −1.23 + 4.82 FG% − 2.59 Opp 3 Pt% + 0.0344 Opp TO
d. Aumenta FG%; disminuye Opp 3 Pt%; aumenta Opp TO
e. 0.638

12. **a.** $R^2 = \dfrac{\text{SCR}}{\text{SCT}} = \dfrac{14\,052.2}{15\,182.9} = 0.926$
b. $R_a^2 = 1 - (1 - R^2)\dfrac{n - 1}{n - p - 1}$
$= 1 - (1 - 0.926)\dfrac{10 - 1}{10 - 2 - 1} = 0.905$
c. Sí; después de ajustar el número de variables independientes del modelo, se ve que 90.5% de la variabilidad en y ha sido explicada.

14. **a.** 0.75 **b.** 0.68

FIGURA D15.5a

```
The regression equation is
Revenue = 88.6 + 1.60 TVAdv

Predictor      Coef    SE Coef       T       p
Constant     88.638      1.582   56.02   0.000
TVAdv        1.6039     0.4778    3.36   0.015

S = 1.215     R-sq = 65.3%     R-sq(adj) = 59.5%

Analysis of Variance

SOURCE            DF       SS        MS        F       p
Regression         1   16.640    16.640    11.27   0.015
Residual Error     6    8.860     1.477
Total              7   25.500
```

FIGURA D15.5b

```
The regression equation is
Revenue = 83.2 + 2.29 TVAdv + 1.30 NewsAdv

Predictor      Coef    SE Coef       T       p
Constant     83.230      1.574   52.88   0.000
TVAdv        2.2902     0.3041    7.53   0.001
NewsAdv      1.3010     0.3207    4.06   0.010

S = 0.6426    R-sq = 91.9%    R-sq(adj) = 88.7%

Analysis of Variance

SOURCE            DF        SS       MS       F       p
Regression         2    23.435   11.718   28.38   0.002
Residual Error     5     2.065    0.413
Total              7    25.500
```

15. a. $R^2 = \dfrac{\text{SCR}}{\text{SCT}} = \dfrac{23.435}{25.5} = 0.919$

$R_a^2 = 1 - (1 - R^2)\dfrac{n - 1}{n - p - 1}$

$= 1 - (1 - 0.919)\dfrac{8 - 1}{8 - 2 - 1} = 0.887$

b. Se prefiere el análisis de regresión múltiple porque tanto R^2 como R_a^2 muestran un mayor porcentaje de variabilidad de y explicada cuando se usan ambas variables independientes

16. a. No, $R^2 = 0.153$

b. Mejor ajuste con regresión múltiple

18. a. $R^2 = 0.564$, $R_a^2 = 0.511$

b. El ajuste no es muy bueno

19. a. $\text{CMR} = \dfrac{\text{SCR}}{p} = \dfrac{6216.375}{2} = 3108.188$

$\text{CME} = \dfrac{\text{SCE}}{n - p - 1} = \dfrac{507.75}{10 - 2 - 1} = 72.536$

b. $F = \dfrac{\text{CMR}}{\text{CME}} = \dfrac{3108.188}{72.536} = 42.85$

De acuerdo con la tabla F (2 grados de libertad en el numerador y 7 en el denominador), el valor-p es menor que 0.01

Usando Excel o Minitab, el valor-p correspondiente a $F =$ 42.85 es 0.0001

Como valor-$p \leq \alpha$, el modelo general es significativo

c. $t = \dfrac{b_1}{s_{b_1}} = \dfrac{0.5906}{0.0813} = 7.26$

valor-p = 0.0002

Como valor-$p \leq \alpha$, β_1 es significativa

d. $t = \dfrac{b_2}{s_{b_2}} = \dfrac{0.4980}{0.0567} = 8.78$

valor-p= 0.0001

Como valor-$p \leq \alpha$, β_2 es significativa

20. a. Significativa; valor-p = 0.000

b. Significativa; valor-p = 0.000

c. Significativa; valor-p = 0.002

22. a. SCE = 4 000, $s^2 = 571.43$, CMR = 6 000

b. Significativa; valor-p = 0.008

23. a. $F = 28.38$

Valor-p = 0.002

Como el valor-$p \leq \alpha$, existe una relación significativa

b. $t = 7.53$

Valor-p = 0.001

Como el valor-$p \leq \alpha$, β_1 es significativa y x_i no debe ser eliminada del modelo

c. $t = 4.06$

Valor-p = 0.010

Como el valor-$p \leq \alpha$, β_2 es significativa y x_2 no debe ser eliminada del modelo

24. a. Relación significativa; valor-p = 0.000

b. HR; Rechazar H_0: $\beta_1 = 0$; valor-p = 0.000

ERA; Rechazar H_0: $\beta_2 = 0$; valor-p = 0.000

26. a. Significativa; valor-p = 0.000

b. Todas significativas; los valor-$p < \alpha = 0.05$

28. a. Usando Minitab, el intervalo de confianza de 95% es de 132.16 a 154.15

b. Usando Minitab, el intervalo de predicción de 95% es de 111.15 a 175.17

29. a. Observe los resultados de Minitab en la figura D15.5b

$\hat{y} = 83.230 + 2.2902(3.5) + 1.3010(1.8) = 93.588$ o $93 588

b. Resultados de Minitab: 92.840 a 94.335 o $92 840 a $94 335

c. Resultados de Minitab: 91.774 a 95.401 o $91.774 a $95 401

30. **a.** 46.758 a 50.646
b. 44.815 a 52.589

32. **a.** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

donde $x_2 = \begin{cases} 0 \text{ si nivel } 1 \\ 1 \text{ si nivel } 2 \end{cases}$

b. $E(y) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1$
c. $E(y) = \beta_0 + \beta_1 x_1 + \beta_2(1) = \beta_0 + \beta_1 x_1 + \beta_2$
d. $\beta_2 = E(y \mid \text{nivel } 2) - E(y \mid \text{nivel } 1)$
β_1 es la variación en $E(y)$ por una variación de una unidad en x_1 cuando x_2 permanece constante

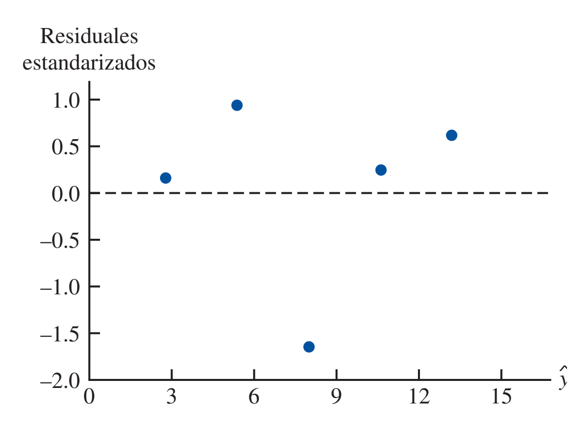
34. **a.** $15 300
b. $\hat{y} = 10.1 - 4.2(2) + 6.8(8) + 15.3(0) = 56.1$
Predicción de las ventas: $56 100
c. $\hat{y} = 10.1 - 4.2(1) + 6.8(3) + 15.3(1) = 41.6$
Predicción de las ventas: $41 600

36. **a.** $\hat{y} = 1.86 + 0.291$ meses $+ 1.10$ tipo $- 0.609$ persona
b. Significativa; valor-$p = 0.002$
c. Persona no es significativa; valor-$p = 0.17$

38. **a.** $\hat{y} = -91.8 + 1.08$ edad $+ 0.252$ presión $+ 8.74$ fumador
b. Significativa; valor-$p = 0.01$
c. El intervalo de predicción del 95% es de 21.35 a 47.18 o una probabilidad de 0.2135 a 0.4718; dejar de fumar y empezar algún tipo de tratamiento para reducir la presión sanguínea

39. **a.** Los resultados de Minitab se presentan en la figura D15.39
b. Minitab proporciona los valores siguientes

x_i	y_i	$\hat{y}_i$	Residual estandarizado
1	3	2.8	0.16
2	7	5.4	0.94
3	5	8.0	−1.65
4	11	10.6	0.24
5	14	13.2	0.62



El punto (3,5) no parece seguir la tendencia del resto de los datos; sin embargo, el valor del residual estandarizado de este punto, −1.65, no es lo suficientemente grande para concluir que (3,5) es una observación atípica

c. Minitab proporciona los valores siguientes:

x_i	y_i	Residual estudentizados eliminados
1	3	0.13
2	7	0.91
3	5	−4.42
4	11	0.19
5	14	0.54

$t_{0.025} = 4.303(n - p - 2 = 5 - 5 - 1 - 2 = 2$ grados de libertad)
Como el residual estudentizado eliminado de (3,5) es $-4.42 < -4.303$ se concluye que la 3ra observación es una observación atípica

40. **a.** $\hat{y} = -53.3 + 3.11x$
b. −1.94, −.012, 1.79, 0.40, −1.90; no

FIGURA D15.39

```
The regression equation is
Y = 0.20 + 2.60 X

Predictor       Coef    SE Coef       T        p
Constant       0.200      2.132    0.09    0.931
X             2.6000     0.6429    4.04    0.027

S = 2.033     R-sq = 84.5%     R-sq(adj) = 79.3%

Analysis of Variance
SOURCE              DF         SS        MS        F        p
Regression           1     67.600    67.600    16.35    0.027
Residual Error       3     12.400     4.133
Total                4     80.000
```

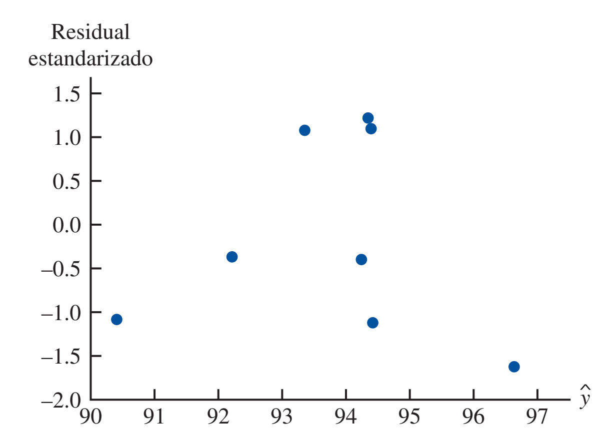
c. 0.38, 0.28, 0.22, 0.20, 0.92; no
d. 0.60, 0.00, 0.26, 0.03, 11.09; sí, la quinta observación

41. a. Los resultados de Minitab se presentan en la figura D15.5; la ecuación de regresión estimada es

Ingreso = 83.2 + 2.29 Publ.Tv + 1.30 Publ.Periód

b. Minitab da los valores siguientes:

$\hat{y}_i$	Residual estandarizado	$\hat{y}_i$	Residual estandarizado
96.63	−1.62	94.39	1.10
90.41	−1.08	94.24	−0.40
94.34	1.22	94.42	−1.12
92.21	−0.37	93.35	1.08



Con relativamente pocas observaciones, es difícil determinar si algunas de las suposiciones con respecto a ϵ se han violado; para este caso, podría analizarse el hecho de que no aparece ningún patrón en el diagrama; alternativamente, puede analizarse la existencia de un patrón curvilíneo en el diagrama.

c. Los valores residuales estandarizados son mayores que −2 y menores que +2; así, usando esta prueba, no hay afloramientos.

Como otra comprobación para afloramientos, utilizamos Minitab para calcular los siguientes residuales studentizados suprimidos:

Observación	Residual studentizado suprimido	Observación	Residual studentizado suprimido
1	−2.11	5	1.13
2	−1.10	6	−0.36
3	1.31	7	−1.16
4	−0.33	8	1.10

$t_{0.025} = 2.776$ ($n - p - 2 = 8 - 2 - 2 = 4$ grados de libertad)

Como ninguno de los residuales studentizados suprimidos es menor de −2.776 o mayor que 2.776, concluimos que no hay afloramientos en los datos

d. Minitab da los valores siguientes:

Observación	h_i	D_i
1	0.63	1.52
2	0.65	0.70
3	0.30	0.22
4	0.23	0.01
5	0.26	0.14
6	0.14	0.01
7	0.66	0.81
8	0.13	0.06

El valor crítico de influencia es

$$\frac{3(p+1)}{n} = \frac{3(2+1)}{8} = 1.125$$

Como ninguno de los valores es mayor que 1.125, se concluye que no hay observaciones influyentes; sin embargo, usando la medida de la distancia de Cook se ve que $D_1 > 1$ (regla práctica del valor crítico); por tanto se concluye que la primera observación es influyente

Conclusión final: la observación 1 es una observación influyente

42. b. Tendencia inusual
c. No hay observaciones atípicas
d. La observación 2 es una observación influyente

44. a. $E(y) = \dfrac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$
b. Estimación de la probabilidad de que un cliente que no tenga tarjeta de crédito de Simmons haga una compra
c. $\hat{g}(x) = -0.9445 + 1.0245x$
d. 0.28 para los clientes que no tienen tarjeta de crédito de Simmons, 0.52 para los clientes que tienen tarjeta de crédito de Simmons
e. Cociente de posibilidades estimado = 2.79

46. a. $E(y) = \dfrac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$
b. $E(y) = \dfrac{e^{-2.6355+0.22018x}}{1 + e^{-2.6355+0.22018x}}$
c. Significativa; valor-p = 0.0002
d. 0.39
e. $1 200
f. Cociente de posibilidades estimado = 1.25

48. a. $E(y) = \dfrac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$
b. $\hat{g}(x) = -2.805 + 1.1492x$
c. 0.86
d. Cociente de posibilidades estimado = 3.16

50. b. 67.39

52. a. $\hat{y} = -1.41 + 0.0235x_1 + 0.00486x_2$
b. Significativa; valor-p = 0.0001
c. $R^2 = 0.937$; $R^2_a = 9.19$; buen ajuste
d. Ambas son significativas

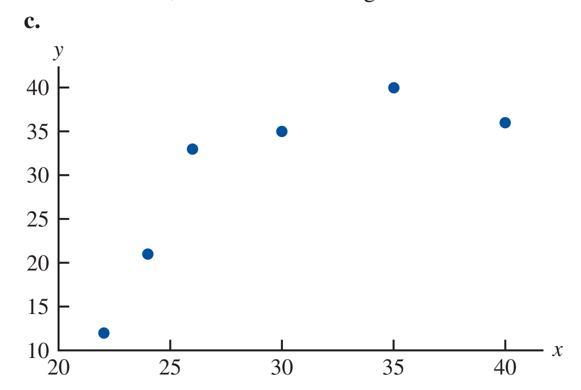
54. a. Marcador = 50.6 + 1.56 RecRes
b. $r^2 = 0.431$; no es una buena forma

c. Puntuación = 33.5 + 1.90 Resist Reces. + 2.61 Acces. Significativa
$R_a^2 = 0.784$; mucho mejor ajuste.

56. a. MPG en ciudad = 24.1 − 2.10 desplazamiento
Significativa; valor-p = 0.000
b. MPG en ciudad = 26.4 − 2.44 desplazamiento − 1.20 tracción 4
c. Significativa; valor-p = 0.016
d. MPG en ciudad = 33.3 – 4.15 desplazamiento – 1.24 tracción + 2.16 ocho cil.
e. Significativa en general e individualmente

Capítulo 16

1. a. En la figura D16.1a se presenta el resultado de Minitab
b. Como el valor-p correspondiente a $F = 6.85$ es 0.059 $> \alpha = 0.05$; la relación no es significativa
c.



El diagrama de dispersión sugiere que una relación curvilínea puede ser la apropiada
d. El resultado de Minitab se muestra en la figura D16.1d
e. Como el valor-p que corresponde a $F = 25.68$ es $0.013 < \alpha = 0.05$, la relación es significativa
f. $\hat{y} = -168.88 + 12.187(25) - 0.17704(25)^2 = 25.145$

2. a. $\hat{y} = 9.32 + 0.424x$; valor-p = 0.117 indica una débil relación entre x y y
b. $\hat{y} = -8.10 + 2.41x - 0.0480x^2$
$R_a^2 = 0.932$; un buen ajuste
c. 20.965

4. a. $\hat{y} = 943 + 8.71x$
b. Significativa; valor-$p = 0.005 < \alpha = 0.01$

5. a. El resultado de Minitab se presenta en la figura D16.5a
b. Como el valor-p que corresponde a $F = 73.15$ es 0.003 $< \alpha = 0.01$, la relación es significativa; puede rechazar H_0: $\beta_1 = \beta_2 = 0$
c. Vea la figura D16.5c

6. b. No, la relación parece ser curvilínea
c. Varios modelos posibles; por ejemplo,
$\hat{y} = 2.90 - 0.185x + 0.00351x^2$

8. a. Parece ser que el modelo de regresión lineal simple no es apropiado
b. Precio = 33829 − 4571 evaluación + 154 evaluación al cuadrado
c. log precio = −10.2 + 10.4 log evaluación
d. Inciso c; se explica un porcentaje mayor de la variación

10. a. Significativa; valor-p = 0.000
b. Significativa; valor-p = 0.000

11. a. SCE = 1 805 − 1 760 = 45

$$F = \frac{\text{CMR}}{\text{CME}} = \left(\frac{1\,760/4}{45/25}\right) = 244.44$$

Como el valor-p = 0.000, la relación es significativa
b. SCE(x_1, x_2, x_3, x_4) = 45
c. SCE(x_2, x_3) = 1805 − 1705 = 100
d. $F = \dfrac{(100 - 45)/2}{1.8} = 15.28 \qquad F_{.05} = 3.39$

Como el valor-p = 0.000, x_1 y x_2 son significativas

FIGURA D16.1a

```
The regression equation is
Y = - 6.8 + 1.23 X

Predictor      Coef    SE Coef        T        p
Constant      -6.77      14.17    -0.48    0.658
X            1.2296     0.4697     2.62    0.059

S = 7.269     R-sq = 63.1%     R-sq(adj) = 53.9%

Analysis of Variance

SOURCE            DF        SS        MS       F        p
Regression         1    362.13    362.13    6.85    0.059
Residual Error     4    211.37     52.84
Total              5    573.50
```

FIGURA D16.1d

```
The regression equation is
Y = - 169 + 12.2 X - 0.177 XSQ

Predictor        Coef    SE Coef        T        p
Constant      -168.88      39.79    -4.74    0.024
X              12.187      2.663     4.58    0.020
XSQ          -0.17704    0.04290    -4.13    0.026

S = 3.248     R-sq = 94.5%     R-sq(adj) = 90.8%

Analysis of Variance

SOURCE            DF       SS       MS        F        p
Regression         2   541.85   270.92    25.68    0.013
Residual Error     3    31.65    10.55
Total              5   573.50
```

FIGURA D16.5a

```
The regression equation is
Y = 433 + 37.4 X -0.383 XSQ

Predictor       Coef    SE Coef        T        p
Constant       432.6      141.2     3.06    0.055
X             37.429      7.807     4.79    0.017
XSQ          -0.3829     0.1036    -3.70    0.034

S = 15.83     R-sq = 98.0%     R-sq(adj) = 96.7%

Analysis of Variance

SOURCE            DF       SS       MS        F        p
Regression         2    36643    18322    73.15    0.003
Residual Error     3      751      250
Total              5    37395
```

FIGURA D16.5c

```
    Fit    Stdev.Fit              95% C.I.              95% P.I.
1302.01      9.93       (1270.41, 1333.61)    (1242.55, 1361.47)
```

12. a. El resultado de Minitab se muestra en la figura D16.12a

b. El resultado de Minitab se muestra en la figura D16.12b

c. $F = \dfrac{[\text{SCE(reducido)} - \text{SCE(completo)}]/(\#\text{ número de términos extra})}{\text{CME(completo)}}$

$$= \frac{(7.2998 - 4.3240)/2}{0.1663} = 8.95$$

El valor-p correspondiente a $F = 8.95$ (2 grados de libertad en el numerador y 26 en el denominador) es 0.001; como el valor-$p < \alpha = 0.05$, la adición de las dos variables independientes es significativa

14. a. $\hat{y} = -111 + 1.32$ edad $+ 0.296$ presión

b. $\hat{y} = -123 + 1.51$ edad $+ 0.448$ presión $+ 8.87$ fumador $- 0.00276$ edad presión

c. Significativa; valor-$p = 0.000$

16. a. Semanas $= -8.9 + 1.51$ edad

b. Semanas $= -0.07 + 1.73$ edad $- 2.7$ manager $- 15.1$ cabeza $- 17.4$ ventas

c. Igual al inciso b

d. Igual al inciso b

e. Semanas $= 13.1 + 1.64$ edad $- 9.76$ casado $- 19.4$ cabeza $- 29.0$ manager $- 19.0$ ventas

FIGURA D16.12a

```
The regression equation is
Scoring Avg. = 46.3 + 14.1 Putting Avg.

Predictor         Coef    SE Coef     T       p
Constant        46.277      6.026  7.68   0.000
Putting Avg.    14.103      3.356  4.20   0.000

S = 0.510596    R-Sq = 38.7%    R-Sq(adj) = 36.5%

Analysis of Variance

SOURCE            DF        SS        MS       F        p
Regression         1    4.6036    4.6036   17.66   0.0000
Residual Error    28    7.2998    0.2607
Total             29   11.9035
```

FIGURA D16.12b

```
The regression equation is
Scoring Avg. = 59.0 - 10.3 Greens in Reg.
  + 11.4 Putting Avg - 1.81 Sand Saves

Predictor           Coef    SE Coef        T        p
Constant          59.022      5.774    10.22    0.000
Greens in Reg.   -10.281      2.877    -3.57    0.001
Putting Avg.      11.413      2.760     4.14    0.000
Sand Saves       -1.8130     0.9210    -1.97    0.060

S = 0.407808     R-Sq = 63.7%     R-Sq(adj) = 59.5%

Analysis of Variance

Source            DF        SS        MS        F        p
Regression         3    7.5795    2.5265    15.19    0.000
Residual Error    26    4.3240    0.1663
Total             29   11.9035
```

18. a. RPG = −4.05 + 27.6 OBP

b. Existe una gran cantidad de modelos que darán un buen ajuste; el modelo de cinco variables identificado usando el procedimiento de regresión por pasos de Minitab con Alfa- to-enter (alfa para ingresar)= 0.10 y Alfa to remove (alfa para eliminar) = 0.10 se presenta a continuación:

RPG = −0.0909 + 32.2 OBP + 0.109 HR − 21.5 AVG + 0.244 3B − 0.0223 BB

20.

x_1	x_2	x_3	Tratamiento
0	0	0	A
1	0	0	B
0	1	0	C
0	0	1	D

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

22. Factor A: $x_1 = 0$ si nivel 1 y 1 si nivel 2
Factor B:

x_2	x_3	Nivel
0	0	1
1	0	2
0	1	3

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1 x_3$

24. a. No es significativa al nivel de signifcancia 0.05; valor-p = 0.093

b. 139

26. Significativos en general; valor-p = 0.029
Individualmente, ninguna de las variables es significativa al nivel de significancia 0.05. Sería útil un tamaño de muestra más grande

28. $d = 1.60$; la prueba no es concluyente

30. **a.** Si ExS denota la interacción entre el coeficiente de gastos y seguridad
Desempeño % = 23.3 + 222 gastos % − 28.9 ExS

b. R-Sq(adj) 0 65.3%; no está mal

c. 25.8 o aproximadamente 26%

32. **a.** AUDELAY = 63.0 + 11.1 INDUS; autocorrelación positiva no significativa

34. Diferencias significativas entre los niveles de confort de los tres tipos de compradores; valor-$p = 0.34$

Capítulo 17

1. **a.**

Artículo	Precio relativo
A	103 = (7.75/7.50)(100)
B	238 = (1500/630)(100)

b. $I_{2006} = \frac{7.75 + 1\,500.00}{7.50 + 630.00}(100) = \frac{1\,507.75}{637.50}(100) = 237$

c. $I_{2006} = \frac{7.75(1\,500) + 1\,500.00(2)}{7.50(1\,500) + 630.00(2)}(100)$

$\frac{14\,625.00}{12\,510.00}(100) = 117$

d. $I_{2006} = \frac{7.75(1\,800) + 1\,500.00(1)}{7.50(1\,800) + 630.00(1)}(100)$

$\frac{15\,450.00}{14\,130.00}(100) = 109$

2. **a.** 32%

b. \$8.14

3. **a.** Precios relativos de A = (6.00/5.45)100 = 110
B = (5.95/5.60)100 = 106
C = (6.20/5.50)100 = 113

b. $I_{2006} = \frac{6.00 + 5.95 + 6.20}{5.45 + 5.60 + 5.50}(100) = 110$

c. $I_{2006} = \frac{6.00(150) + 5.95(200) + 6.20(120)}{5.45(150) + 5.60(200) + 5.50(120)}(100)$
$= 109$

incremento de 9% en el periodo de dos años

4. $I_{2006} = 122$

6.

Artículo	Precio Relativo	Periodo Base Precio	Periodo Base Uso	Peso	Peso relativo ponderado
A	150	22.00	20	440	66 000
B	90	5.00	50	250	22 500
C	120	14.00	40	560	67 200
			Totales	1250	155 700

$I = \frac{155\,700}{1\,250} = 125$

7. **a.** Precios relativos de A 5 (3.95/2.50)100 5 158
B = (9.90/8.75)100 = 113
C = (0.95/.99)100 = 96

b.

Artículo	Precio relativo	Periodo base	Calidad	Peso $P_{i0}Q_i$	Peso relativo ponderado
A	158	2.50	25	62.5	9 875
B	113	8.75	15	131.3	14 837
C	96	0.99	60	59.4	5 702
			Totales	253.2	30 414

$I = \frac{30\,414}{253.2} = 120$

El costo de las materias primas aumentó 20%

8. $I = 105$; el portafolio aumentó 5%

10. **a.** Salarios deflactados de 1996: $\frac{\$11.86}{154.9}(100) = \$7.66$

Salarios deflactados de 2006: $\frac{\$16.47}{198.7}(100) = \$8.29$

b. $\frac{16.47}{11.86}(100) = 138.9$; el cambio porcentual en los salarios reales es un incremento de 8.2%

c. $\frac{8.29}{7.66}(100) = 108.2$; el cambio porcentual en los salarios reales es un incremento de 8.2%

12. **a.** 2 420, 2 449, 2 242
Los pedidos de la industria aumentaron ligeramente en términos de dólares constantes

b. 3 032, 3 057, 2 822

c. PPI

14. $I = \frac{300(18.00) + 400(4.90) + 850(15.00)}{350(18.00) + 220(4.90) + 730(15.00)}(100)$

$= \frac{20\,110}{18\,328}(100) = 110$

15. $I = \frac{95(1\,200) + 75(1\,800) + 50(2\,000) + 70(1\,500)}{120(1\,200) + 86(1\,800) + 35(2\,000) + 60(1\,500)}(100)$
$= 99$

Las cantidades disminuyeron ligeramente

16. $I = 83$

18. **a.** 151, 197, 143, 178

b. $I = 170$

20. $I_{enero} = 73.5$, $I_{marzo} = 70.1$

22. $I = 86.2$

24. \$36 082; \$32 528; \$27 913; \$34 387; \$40 551; \$42 651; \$46 458; \$56 324

26. $I = 143\,143$; la cantidad aumentó 43%

Capítulo 18

1. a.

Semana	Valor de la serie de tiempos	Pronóstico	Error de pronóstico	Cuadrado del error de pronóstico
1	8			
2	13			
3	15			
4	17	12	5	25
5	16	15	1	1
6	9	16	−7	49
			Total	75

El pronóstico para la semana 7 es (17 + 16 + 9)/3 = 14

b. CME = 75/3 = 25

c.

Semana (t)	Valor de la serie de tiempos (Y_t)	Pronóstico F_t	Error de pronóstico $Y_t - F_t$	Cuadrado del error de pronóstico $(Y_t - F_t)^2$
1	8			
2	13	8.00	5.00	25.00
3	15	9.00	6.00	36.00
4	17	10.20	6.80	46.24
5	16	11.56	4.44	19.71
6	9	12.45	−3.45	11.90
Total				138.85

El pronóstico para la semana 7 es 2(9) + 0.8(12.45) = 11.76

d. Para $\alpha = 0.2$ pronóstico suavizado exponencial

$$\text{CME} = \frac{138.85}{5} = 27.77$$

Como el promedio móvil de tres semanas tienen un CME menor, parece ser el que proporciona una mejor predicción

e.

Semana (t)	Valor de la serie de tiempos (Y_t)	Pronóstico F_t	Error de pronóstico $Y_t - F_t$	Cuadrado del error de pronóstico $(Y_t - F_t)^2$
1	8			
2	13	8.0	5.0	25.00
3	15	10.0	5.0	25.00
4	17	12.0	5.0	25.00
5	16	14.0	2.0	4.00
6	9	14.8	−5.8	33.64
			Total	112.64

$$\text{CME} = \frac{112.64}{5} = 22.53$$

Una constante de suavizamiento de 0.4 es la que parece proporcionar los mejores pronósticos; el pronóstico para la semana 7 usando $\alpha = 0.4$ es 0.4(9) + 0.6(14.8) = 12.48

2. a.

Semana	Cuatro semanas	Cinco semanas
10	19.00	18.80
11	20.00	19.20
12	18.75	19.00

b. 9.65, 7.41

c. Cinco semanas

4. Semanas 10, 11, y 12: 18.48, 18.63, 18.27
CME = 9.25; $\alpha = 0.2$ es mejor

6. a. CME (tres meses) = 1.24
CME ($\alpha = 0.2$) = 3.55
Usar promedios móviles de tres meses

b. 83.3

8. a.

Mes	Valor de la serie de tiempo	Pronóstico con un promedio móvil de tres meses	(Error)^2	$\alpha = 0.2$ Pronóstico	(Error)^2
1	240				
2	350		240.00	12 100.00	
3	230		262.00	1 024.00	
4	260	273.33	177.69	255.60	19.36
5	280	280.00	0.00	256.48	553.19
6	320	256.67	4 010.69	261.18	3 459.79
7	220	286.67	4 444.89	272.95	2 803.70
8	310	273.33	1 344.69	262.36	2 269.57
9	240	283.33	1 877.49	271.89	1 016.97
10	310	256.67	2 844.09	265.51	1 979.36
11	240	286.67	2 178.09	274.41	1 184.05
12	230	263.33	1 110.89	267.53	1 408.50
Totales			17 988.52		27 818.49

MCE (tres meses) = 17 988.52/9 = 1 998.72

MCE ($\alpha = 0.2$) = 27 818.49/11 = 2528.95

De acuerdo con los valores de CME anteriores, el promedio móvil de tres meses parece ser mejor; sin embargo, el suavizamiento exponencial se ve perjudicado por incluir el mes 2, que era complicado para cualquier método de pronóstico

Usando sólo los errores de los meses 4 – 12, el CME del suavizamiento exponencial se modifica a

MCE($\alpha = 0.2$) = 14 694.49/9 = 1632.72

Por tanto el suavizamiento exponencial fue mejor cuando se consideraron los meses 4 – 12

b. Empleando suavizamiento exponencial

$$F_{13} = \alpha Y_{12} + (1 - \alpha)F_{12}$$
$$= 0.20(230) + 0.80(267.53) = 260$$

10. c. Use $\alpha = 0.3$; $F_{11} = 7.57$

12. $\Sigma t = 15$, $\Sigma t^2 = 55$, $\Sigma Y_t = 55$, $\Sigma tY_t = 186$

$$b_1 = \frac{\Sigma tY_t - (\Sigma t\,\Sigma Y_t)/n}{\Sigma t^2 - (\Sigma t)^2/n}$$

$$= \frac{186 - (15)(55)/5}{55 - (15)^2/5} = 2.1$$

$$b_0 = \bar{Y} - b_1\bar{t} = 11 - 2.1(3) = 4.7$$

$$T_t = 4.7 + 2.1t$$

$$T_6 = 4.7 + 2.1(6) = 17.3$$

14. $\Sigma t = 21, \Sigma t^2 = 91, \Sigma Y_t = 117.1, \Sigma tY_t = 403.7$

$$b_1 = \frac{\Sigma tY_t - (\Sigma t\, \Sigma Y_t)/n}{\Sigma t^2 - (\Sigma t)^2/n}$$

$$= \frac{403.7 - (21)(117.1)/6}{91 - (21)^2/6} = -0.3514$$

$$b_0 = \bar{Y} - b_1\bar{t} = 19.5167 - (-0.3514)(3.5) = 20.7466$$

$$T_t = 20.7466 - 0.3514t$$

La matrícula parece estar disminuyendo, la disminución es de alrededor de 351 estudiantes por año

16. Considerar una tendencia no lineal

18. **a.** Rurales: $T_t = -4 + 5.2\,t$
Urbanos: $T_t = 2.3 + 6.9t$
Suburbanos: $T_t = 1.4 + 7.4t$
b. 5.2%, 6.9%, 7.4%
c. 27.2%, 43.7%, 45.8%

20. **a.** $T_t = 1997.6 + 397.545t$
b. $T_{11} = 6371, T_{12} = 6768$

22. **a.**

Año	Trimestre	Y_t	Promedio móvil de cuatro trimestres	Promedio móvil centrado
1	1	4		
	2	2		
			3.50	
	3	3		3.750
			4.00	
	4	5		4.125
			4.25	
2	1	6		4.500
			4.75	
	2	3		5.000
			5.25	
	3	5		5.375
			5.50	
	4	7		5.875
			6.25	
3	1	7		6.375
			6.50	
	2	6		6.625
			6.75	
	3	6		
	4	8		

b.

Año	Trimestre	Y_t	Promedio móvil centrado	Componente estacional irregular
1	1	4		
	2	2		
	3	3	3.750	0.8000
	4	5	4.125	1.2121
2	1	6	4.500	1.3333
	2	3	5.000	0.6000
	3	5	5.375	0.9302
	4	7	5.875	1.1915
3	1	7	6.375	1.0980
	2	6	6.625	0.9057
	3	6		
	4	8		

Trimestre	Valor del componente estacional irregular	Índice estacional
1	1.3333, 1.0980	1.2157
2	0.6000, 0.9057	0.7529
3	0.8000, 0.9302	0.8651
4	1.2121, 1.1915	1.2018
	Total	4.0355

Ajuste del índice estacional $= \frac{4}{4.0355} = 0.9912$

Trimestre	Índice estacional ajustado
1	1.2050
2	0.7463
3	0.8575
4	1.1912

24. Índices estacionales ajustados: 0.707, 0.777, 0.827, 0.966, 1.016, 1.305, 1.494, 1.225, 0.976, 0.986, 0.936, 0.787
Nota: ajuste = 0.996

26. **a.** Sí
b. 12–4: 166 761.13
4–8: 146 052.99

28. **a.** 0.3 es mejor
b. 18.41

30. 20.26

32. **a.** $\alpha = 0.5$
b. $T_t = 244.778 + 22.088t$
c. Proyección de tendencia: CME más pequeño

34. $T_8 = 252.28, T_9 = 259.10$

36. **a.** Sí
b. $T_t = -5 + 15t$

38. **a.** Una tendencia lineal parece ser adecuada
b. $T_t = 6.4564 + 0.5345t$

c. 0.5345 millones
d. 12.87 millones

40. b. Índice estacional ajustado: 0.899, 1.362, 1.118, 0.621
Nota: ajuste = 1. 0101
c. Trimestre 2; parece razonable

42. a. $T_t = 6.329 + 1.055t$
b. 36.92, 37.98, 39.03, 40.09
c. 33.23, 51.65, 43.71, 24.86

Capítulo 19

1. Probabilidades binomiales para $n = 10$, $p = 0.50$

x	**Probabilidad**	*x*	**Probabilidad**
0	0.0010	6	0.2051
1	0.0098	7	0.1172
2	0.0439	8	0.0439
3	0.1172	9	0.0098
4	0.2051	10	0.0010
5	0.2461		

Cantidad de signos positivos = 7

$P(x \geq 7) = P(7) + P(8) + P(9) + P(10)$
$= 0.1172 + 0.0439 + 0.0098 + 0.0010$
$= 0.1719$

Valor-$p = 2(0.1719) = 0.3438$

Valor-$p > 0.05$; no rechazar H_0

No existe indicación de diferencia

2. $n = 27$ casos en los que se obtuvo un valor diferente a 150
Use la aproximación normal con $\mu = np = 0.5(27) = 13.5$ y $\sigma = \sqrt{.25n} = \sqrt{.25(27)} = 2.6$
Use $x = 22$ como número de signos "más" para obtener el estadístico de prueba siguiente:

$$z = \frac{x - \mu}{\sigma} = \frac{22 - 13.5}{2.6} = 3.27$$

En las tablas mayor valor $z = 3.09$
Área en la cola = 1.000 – 0.9990 = 0.001
Para $z = 3.27$, el valor-p es menor que 0.001
Valor-$p \leq$ 0.01; rechazar H_0 y concluir que la mediana > 150

4. Se necesita determinar la cantidad de respuestas "mejor" y la cantidad de respuestas "peor", su suma es el tamaño de la muestra en el estudio

$$n = 0.34(1\,253) + 0.29(1\,253) = 789.4$$

Use la prueba para muestras grandes y la distribución normal; el valor de $n = 789.4$ no necesita ser entero.

$$\text{Use} \quad \begin{array}{l} \mu = 0.5n = 0.5(789.4) = 394.7 \\ \sigma = \sqrt{0.25n} = \sqrt{0.25(789.4)} = 14.05 \end{array}$$

Sea p la proporción de adultos que creen que sus hijos tendrán un mejor futuro

H_0: $p \leq 0.50$

H_a: $p > 0.50$

$$x = 0.34(1253) = 426.0$$

$$z = \frac{x - \mu}{\sigma} = \frac{426.0 - 394.7}{14.05} = 2.23$$

Valor-p = 1.0000 – 0.9871 = 0.0129

Rechazar H_0 y concluir que hay más adultos que piensan que sus hijos tendrán un mejor futuro

6. $z = 2.32$
Valor-$p = 0.0204$
Rechazar H_0

8. $z = 3.76$
Valor-$p \approx 0$
Rechazar H_0

10. $z = 1.27$
Valor-$p = 0.2040$
No rechazar H_0

12. H_0: Las poblaciones son idénticas
H_a: Las poblaciones no son idénticas

Aditivo 1	**Aditivo 2**	**Diferencia**	**Valor absoluto**	**Rango**	**Rango con signo**
20.12	18.05	2.07	2.07	9	+9
23.56	21.77	1.79	1.79	7	+7
22.03	22.57	−0.54	0.54	3	−3
19.15	17.06	2.09	2.09	10	+10
21.23	21.22	0.01	0.01	1	+1
24.77	23.80	0.97	0.97	4	+4
16.16	17.20	−1.04	1.04	5	−5
18.55	14.98	3.57	3.57	12	+12
21.87	20.03	1.84	1.84	8	+8
24.23	21.15	3.08	3.08	11	+11
23.21	22.78	0.43	0.43	2	+2
25.02	23.70	1.32	1.32	6	+6
					$T = 62$

$$\mu_T = 0$$

$$\sigma_T = \sqrt{\frac{n(n + 1)(2n + 1)}{6}} = \sqrt{\frac{12(13)(25)}{6}} = 25.5$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{62 - 0}{25.5} = 2.43$$

Valor-p = 2(1.0000 – 0.9925) = 0.0150

Rechazar H_0 y concluir que hay diferencia significativa entre los aditivos

13.

Sin relajante	**Con relajante**	**Diferencia**	**Rango de las diferencias absolutas**	**Rangos con signo**
15	10	5	9	+9
12	10	2	3	+3
22	12	10	10	+10
8	11	−3	6.5	−6.5
10	9	1	1	+1
7	5	2	3	+3
8	10	−2	3	−3
10	7	3	6.5	+6.5
14	11	3	6.5	+6.5
9	6	3	6.5	+6.5
				$T = 36$

$\mu_T = 0$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} = \sqrt{\frac{10(11)(21)}{6}} = 19.62$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{36}{19.62} = 1.83$$

Valor-p = 1.0000 – 0.9664 = 0.0336

Rechazar H_0 y concluir que existe diferencia significativa a favor del relajante

14. $z = 2.29$
Valor-p = 0.0220
Rechazar H_0

16. $z = -1.48$
Valor-p = 0.1388
Rechazar H_0

18. Ordenar por rango las muestras combinadas (juntas) y hallar la suma de los rangos de cada muestra; ésta es una prueba con una muestra pequeña porque $n_1 = 7$ y $n_2 = 9$

Aditivo 1		Aditivo 2	
MPG	Rango	MPG	Rango
17.3	2	18.7	8.5
18.4	6	17.8	4
19.1	10	21.3	15
16.7	1	21.0	14
18.2	5	22.1	16
18.6	7	18.7	8.5
17.5	3	19.8	11
	34	20.7	13
		20.2	12
			102

$T = 34$

Con $\alpha = 0.05$, $n_1 = 7$, y $n_2 = 9$

$$T_L = 41 \text{ y } T_U = 7(7 + 9 + 1 - 41 = 98$$

Como $T = 34 < 41$, rechazar H_0 y concluir que hay diferencia significativa en el rendimiento de la gasolina

19. a.

Contador público	Rango	Planificador financiero	Rango
45.2	5	44.0	2
53.8	19	44.2	3
51.3	16	48.1	10
53.2	18	50.9	15
49.2	13	46.9	8.5
50.0	14	48.6	11
45.9	6	44.7	4
54.5	20	48.9	12
52.0	17	46.8	7
46.9	8.5	43.9	1
	136.5		73.5

$$\mu_T = \frac{1}{2}n_1(n_1 + n_2 + 1) = \frac{1}{2}(10)(10 + 10 + 1) = 105$$

$$\sigma_T = \sqrt{\frac{1}{12}n_1n_2(n_1 + n_2 + 1)} = \sqrt{\frac{1}{12}(10)(10)(10 + 10 + 1)}$$
$$= 13.23$$

$T = 136.5$

$$z = \frac{136.5 - 105}{13.23} = 2.38$$

Valor-p = 2(1.0000 − 0.9913) = 0.0174

Rechazar H_0 y concluir que los salarios difieren considerablemente entre las dos profesiones

b. Contador público $50 200
Planificador financiero $46 700

20. a. Mujeres 49.9, Hombres 35.4
b. $T = 36$, $T_L = 37$
Rechazar H_0

22. $z = 2.77$
Valor-p = 0.0056
Rechazar H_0

24. $z = -0.25$
Valor-p = 0.8026
Rechazar H_0

26. Calificaciones:

Producto A	Producto B	Producto C
4	11	7
8	14	2
10	15	1
3	12	6
9	13	5
34	65	21

$$W = \frac{12}{(15)(16)}\left[\frac{(34)^2}{5} + \frac{(65)^2}{5} + \frac{(21)^2}{5}\right] - 3(15 + 1)$$
$$= 58.22 - 48 = 10.22 \quad (df = 2)$$

Valor-p está entre 0.005 y 0.01

Rechazar H_0 y concluir que las calificaciones dadas a los productos difieren

28. Clasificaciones:

Natación	Tenis	Bicicleta
8	9	5
4	14	1
11	13	3
6	10	7
12	15	2
41	61	18

$$W = \frac{12}{15(15 + 1)}\left[\frac{41^2}{5} + \frac{61^2}{5} + \frac{18^2}{5}\right] - 3(15 + 1)$$
$$= 9.26 \quad (df = 2)$$

El valor-p está entre 0.005 y 0.01

Rechazar H_0 y concluir que hay diferencia entre las actividades

30. $W = 8.03$; $gl = 3$
El valor-p está entre 0.025 y 0.05
Rechazar H_0

32. a. $\Sigma d_i^2 = 52$

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)} = 1 - \frac{6(52)}{10(99)} = 0.68$$

b.

$$\sigma_{r_s} = \sqrt{\frac{1}{n-1}} = \sqrt{\frac{1}{9}} = 0.33$$

$$z = \frac{r_s - 0}{\sigma_r} = \frac{0.68}{0.33} = 2.05$$

El valor-p = 2(1.0000 − 0.9798) = 0.0404

Rechazar H_0 y concluir que existe una correlación de rangos significativa

34. $\Sigma d_i^2 = 250$

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6(250)}{11(120)} = -0.136$$

$$\sigma_{r_s} = \sqrt{\frac{1}{n-1}} = \sqrt{\frac{1}{10}} = 0.32$$

$$z = \frac{r_s - 0}{\sigma_{r_s}} = \frac{-0.136}{0.32} = -0.43$$

Valor-p = 2(0.3336) = 0.6672

No rechazar H_0; no se puede concluir que haya una relación significativa entre los rangos

36. $r_s = -0.71$, $z = -2.13$
Valor-p = 0.0332
Rechazar H_0

38. $z = -3.17$
Valor-p = 0.002
Rechazar H_0

40. $z = -2.59$
Valor-p = 0.0096
Rechazar H_0

42. $z = -2.97$
Valor-p = 0.003
Rechazar H_0

44. $W = 12.61$; $gl = 2$
El valor-p está entre 0.01 y 0.025
Rechazar H_0

46. $r_s = 0.76$, $z = 2.83$
Valor-p = 0.0046
Rechazar H_0

Capítulo 20

2. a. 5.42
b. UCL = 6.09, LCL = 4.75

4. *Carta R:*
UCL = $\bar{R}D_4$ = 1.6(1.864) = 2.98
LCL = $\bar{R}D_3$ = 1.6(.136) = 0.22
Carta $\bar{x}$:
UCL = $\bar{\bar{x}} + A_2\bar{R}$ = 28.5 + 0.373(1.6) = 29.10
LCL = $\bar{\bar{x}} - A_2\bar{R}$ = 28.5 − 0.373(1.6) = 27.90

6. 20.01, 0.082

8. a. 0.0470
b. UCL = 0.0989, LCL = −0.0049 (use LCL = 0)
c. $\bar{p}$ = 0.08; en control
d. UCL = 14.826, LCL = -0.726 (use LCL = 0)
El proceso está fuera de control si hay más de 14 defectuosos
e. En control con 12 defectuosos
f. Carta *np*

10. $f(x) = \frac{n!}{x!(n-x)!} p^x(1-p)^{n-x}$

Si p = 0.02, la probabilidad de aceptar el lote es

$$f(0) = \frac{25}{0\,(25-0)}(0.02)^0(1-0.02)^{25} = 0.6035$$

Si p = 0.06, la probabilidad de aceptar el lote es

$$f(0) = \frac{25}{0\,(25-0)}(0.06)^0(1-0.06)^{25} = 0.2129$$

12. p_0 = 0.02; riesgo del productor = 0.0599
p_0 = 0.06; riesgo del productor = 0.3396
El riesgo del productor disminuye a medida que aumenta el criterio de aceptación c

14. $n = 20$, $c = 3$

16. a. 95.4
b. UCL = 96.07, LCL = 94.73
c. No

18.

	Carta *R*	Carta $\bar{x}$
UCL	4.23	6.57
LCL	0	4.27

Estimación de la desviación estándar 0.86

20.

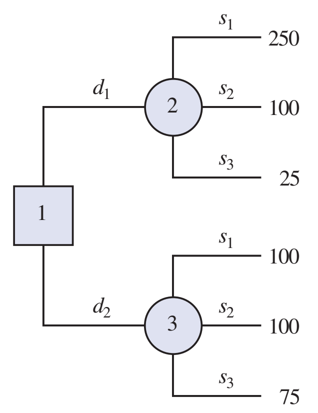
	Carta *R*	Carta $\bar{x}$
UCL	0.1121	3.112
LCL	0	3.051

22. a. UCL = 0.0817, LCL = −0.0017 (use LCL = 0)

24. a. 0.03
b. β = 0.0802

Capítulo 21

1. a.



b. VE(d_1) = 0.65(250) + 0.15(100) + 0.20(25) = 182.5
VE(d_2) = 0.65(100) + 0.15(100) + 0.20(75) = 95
La decisión óptima es d_1

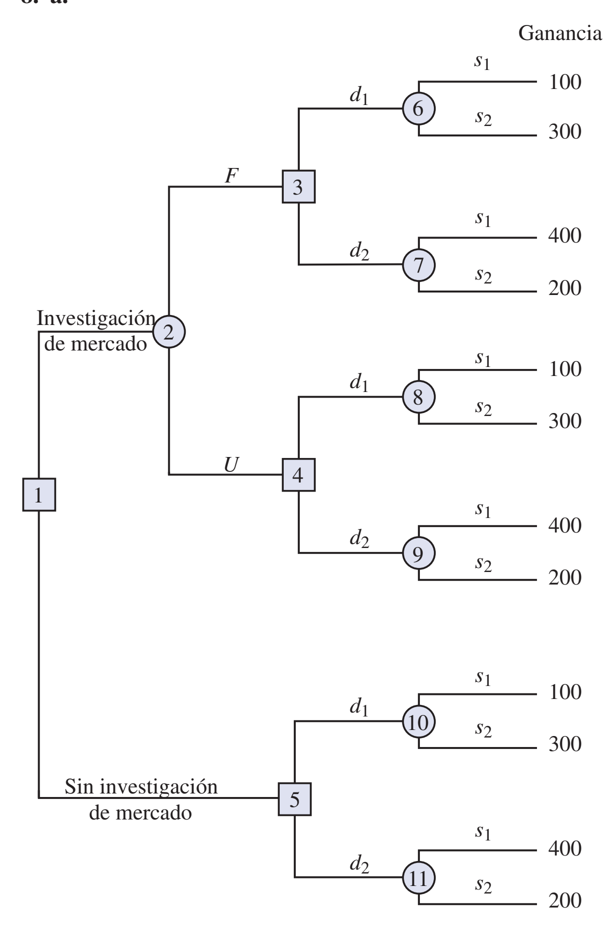
2. a. d_1; VE(d_1) = 11.3
b. d_4; VE(d_4) = 9.5

3. a. VE(propio personal) = 0.2(650) + 0.5(650) + 0.3(600) = 635
VE(empresa externa) = 0.2(900) + 0.5(600) + 0.3(300) = 570
VE(combinación) = 0.2(800) + 0.5(650) + 0.3(500) = 635
Decisión óptima: contratar una empresa externa con un costo esperado de $570 000
b. VEcIP = 0.2(650) + 0.5(600) + 0.3(300) = 520
VEIP = |520 − 570| = 50, o $50 000

4. b. Precio bajo; VE = 565
c. Predio normal; VE = 670

6. c. Sólo Chardonay; VE = 42.5
d. Las dos uvas; VE = 46.4
e. Las dos uvas; VE = 39.6

8. a.



b. VE (nodo 6) = 0.57(100) + 0.43(300) = 186
VE (nodo 7) = 0.57(400) + 0.43(200) = 314
VE (nodo 8) = 0.18(100) + 0.82(300) = 264
VE (nodo 9) = 0.18(400) + 0.82(200) = 236
VE (nodo 10) = 0.40(100) + 0.60(300) = 220
VE (nodo 11) = 0.40(400) + 0.60(200) = 280

VE (nodo 3) = Max(186,314) = 314 d_2
VE (nodo 4) = Max(264,236) = 264 d_1
VE (nodo 5) = Max(220,280) = 280 d_2

VE (nodo 2) = 0.56(314) + 0.44(264) = 292
VE (nodo 1) = Max(292,280) = 292

∴ Investigación de mercado
Si es favorable, decisión d_2
Si es desfavorable, decisión d_1

10. a. 5 000 − 200 − 2 000 − 150 = 2 650
3 000 − 200 − 2 000 − 150 = 650

b. Valores esperados en los nodos:
8: 2 350 5: 2 350 9: 1 100
6: 1 150 10: 2 000 7: 2 000
4: 1 870 3: 2 000 2: 1 560
1: 1 560
c. El costo tendría que disminuir por lo menos a $130 000

12. b. d_1, 1 250
c. 1 700
d. If N, d_1
If U, d_2; 1 666

14.

Situación	$P(s_j)$	$P(I \mid s_j)$	$P(I \cap s_j)$	$P(s_j \mid I)$
s_1	.2	0.10	0.020	0.1905
s_2	.5	0.05	0.025	0.2381
s_3	.3	0.20	0.060	0.5714
	1.0		$P(I)$ = 0.105	1.0000

16. a. 0.695, 0.215, 0.090
0.98, 0.02
0.79, 0.21
0.00, 1.00
c. Si D, autopista
Si N, autopista
Si L, avenida Queen
26.6 minutos

Capítulo 22

1. a. $\bar{x} = 215$ es una estimación de la media poblacional

b. $s_{\bar{x}} = \dfrac{20}{\sqrt{50}}\sqrt{\dfrac{800-50}{800}} = 2.7386$

c. $215 \pm 2(2.7386)$ o 209.5228 a 220.4772

2. a. 30 000
b. 320
c. 29 360 a 30 640

4. 73

5. a. $\bar{x} = 149\,670$ y $s = 73\,420$

$$s_{\bar{x}} = \sqrt{\left(\frac{771-50}{771}\right)\frac{73\,420}{\sqrt{50}}} = 10\,040.83$$

Intervalo de confianza de aproximadamente 95%:

149 760 ± 2(10 040.83)

o

\$129 588.34 a \$169 751.66

b. $\hat{X} = N_{\bar{x}} = 771(149\,670) = 115\,395\,570$

$s_{\hat{X}} = Ns_{\bar{x}} = 771(10\,040.83) = 7\,741\,479.93$

Intervalo de confianza de aproximadamente 95%:

115 395 570 ± 2(7 741 479.93)

o

\$99 912 810.14 a \$130 878 729.86

c. $\bar{p} = {}^{18}\!/_{50} = 0.36$ y

$$s_{\bar{p}} = \sqrt{\left(\frac{771-50}{771}\right)\frac{(0.36)(0.64)}{49}} = 0.0663$$

Intervalo de confianza de aproximadamente 95%:

0.36 ± 2(0.0663)

o

0.2274 a 0.4926

Este intervalo es bastante amplio; los tamaños de la muestra deben ser grandes para obtener intervalos de confianza para la proporción poblacional que sean estrechos.

6. 337

7. a. Estrato 1: $\bar{x}_1 = 138$
Estrato 2: $\bar{x}_2 = 103$
Estrato 3: $\bar{x}_3 = 210$

b. Estrato 1

$$\bar{x}_1 = 138;\ s_{\bar{x}_1} = \left(\frac{30}{\sqrt{20}}\right)\sqrt{\frac{200-20}{200}} = 6.3640$$

Intervalo de confianza de aproximadamente 95%:

138 ± 2(6.3640)

o 125.272 a 150.728

Estrato 2

$$\bar{x}_2 = 103;\ s_{\bar{x}_2} = \left(\frac{25}{\sqrt{30}}\right)\sqrt{\frac{250-30}{250}} = 4.2817$$

Intervalo de confianza de aproximadamente 95%:

103 ± 2(4.2817)

o 99.4366 a 111.5634

Estrato 3

$$\bar{x}_3 = 210;\ s_{\bar{x}_3} = \left(\frac{50}{\sqrt{25}}\right)\sqrt{\frac{100-25}{100}} = 8.6603$$

Intervalo de confianza de aproximadamente 95%:

210 ± 2(8.6603)

o 192.6794 a 227.3206

c. $\bar{x}_{st} = \left(\dfrac{200}{550}\right)138 + \left(\dfrac{250}{550}\right)103 + \left(\dfrac{100}{550}\right)210$

$= 50.1818 + 46.8182 + 38.1818 = 135.1818$

$$s_{\bar{x}_{st}} = \sqrt{\left(\frac{1}{(550)^2}\right)\left(200(180)\frac{(30)^2}{20} + 250(220)\frac{(25)^2}{30} + 100(75)\frac{(50)^2}{25}\right)}$$

$$= \sqrt{\left(\frac{1}{(550)^2}\right)3\,515\,833.3} = 3.4092$$

Intervalo de confianza de aproximadamente 95%:

135.1818 ± 2(3.4092)

o 128.3634 a 142.0002

8. a. Estrato 1: 27 600
Estrato 2: 25 750
Estrato 3: 21 000
b. 74 350
c. 70 599.88 a 78 100.12

10. a. $n = 93,\ n_1 = 30,\ n_2 = 30,\ n_3 = 33$
b. $n = 306,\ n_1 = 98,\ n_2 = 98$
$n_3 = 109$
c. $n = 275,\ n_1 = 88,\ n_2 = 88,\ n_3 = 98$

12. a. \$3 617 000
b. \$1 122 265
c. \$41 066 a \$56 499
d. \$9 568 261 a \$13 164 197

14. a. $\bar{x}_c = \dfrac{\Sigma x_i}{\Sigma M_i} = \dfrac{750}{50} = 15$

$\hat{X} = M\bar{x}_c = 300(15) = 4\,500$

$\bar{p}_c = \dfrac{\Sigma a_i}{\Sigma M_i} = \dfrac{15}{50} = .30$

b.
$$\begin{aligned}\Sigma(x_i - \bar{x}_c M_i)^2 &= [95 - 15(7)]^2 + [325 - 15(18)]^2 \\ &\quad + [190 - 15(15)]^2 + [140 - 15(10)]^2 \\ &= (-10)^2 + (55)^2 + (-35)^2 + (-10)^2 \\ &= 4\,450\end{aligned}$$

$$s_{\bar{x}_c} = \sqrt{\left(\frac{25 - 4}{(25)(4)(12)^2}\right)\left(\frac{4450}{3}\right)} = 1.4708$$

$$s_{\hat{X}} = Ms_{\bar{x}_c} = 300(1.4708) = 441.24$$

$$\begin{aligned}\Sigma(a_i - \bar{p}_c M_i)^2 &= [1 - 0.3(7)]^2 + [6 - 0.3(18)]^2 \\ &\quad + [6 - 0.3(15)]^2 + [2 - 0.3(10)]^2 \\ &= (-1.1)^2 + (0.6)^2 + (1.5)^2 + (-1)^2 \\ &= 4.82\end{aligned}$$

$$s_{\bar{p}_c} = \sqrt{\left(\frac{25 - 4}{(25)(4)(12)^2}\right)\left(\frac{4.82}{3}\right)} = 0.0484$$

c. Intervalo de confianza de aproximadamente 95% para la media poblacional:
15 ± 2(1.4708)
o 12.584 a 17.9416

d. Intervalo de confianza de aproximadamente 95% para el total poblacional:
4 500 ± 2(441.24)
o 3 617.52 a 5 382.48

e. Intervalo de confianza de aproximadamente 95% para la proporción poblacional:
0.30 ± 2(0.0484)
o 0.2032 a 0.3968

16. **a.** 40
b. 0.70
c. 35.8634 a 44.1366
d. 0.5234 a 0.8766

18. **a.** 0.1488 a 0.2312
b. 0.2615 a 0.3585
c. 0.1306 a 0.2094

20. **a.** $22 790 a $23 610
b. $68 370 366 a $70 829 634
c. 0.6692 a 0.7908

22. **a.** 431
b. 0.2175 a 0.3983
c. 0.6230 a 0.8002
d. 996

24. **a.** 75.275
b. 0.198 a 0.502
c. 1 680

Apéndice E: Uso de las funciones de Excel

Excel proporciona una gran cantidad de funciones para el manejo de datos y para el análisis estadístico. Si se sabe cuál es la función que se necesita y cómo usarla, basta ingresar la función en la hoja de cálculo adecuada. Si no se sabe cuál es la función adecuada para la realización de una tarea o no se está seguro de cómo usar una función determinada, Excel puede proporcionar la ayuda necesaria.

Encontrar la función adecuada de Excel

*En las versiones anteriores de Excel, el cuadro de diálogo **Paste function** es lo mismo que el cuadro de diálogo **Insertar función** en Excel 2003.*

Para identificar las funciones disponibles en Excel, seleccione el menú **Insertar** y después, en la lista de opciones, elija **Función.** Otra alternativa es seleccionar, en la barra de fórmulas, el botón f_x. Con cualquiera de los dos métodos aparece el cuadro de diálogo **Insertar función** que se muestra en la figura 1.

El cuadro **Buscar una función** que aparece en la parte superior del cuadro de diálogo Insertar función permite dar una pequeña descripción de lo que se quiere hacer. Una vez hecho esto, se hace clic en **Ir,** Excel buscará y mostrará en el cuadro **Seleccionar una función**, las funciones que puedan realizar la tarea deseada. Sin embargo, con frecuencia se desea echar un vistazo a todas las categorías de funciones para ver con cuáles se cuenta. En este caso puede ser de ayuda el cuadro **O seleccionar una categoría**. Este cuadro tiene una lista desplegable que contiene las diferentes categorías de funciones que proporciona Excel. En la figura 1, la categoría seleccionada es **Estadísticas**. Las funciones estadísticas de Excel aparecen en orden alfabético en el cuadro Seleccione una función. Como se ve, AVEDEV es la primera función de la lista, seguida por la función AVERAGE, etcétera.

FIGURA 1 CUADRO DE DIÁLOGO INSERTAR FUNCIÓN

En la figura 1, la función AVEDEV aparece sombreada, lo que indica que es la función seleccionada en ese momento. Debajo del cuadro Seleccione una función, aparece la sintaxis correcta de la función y una breve descripción de la misma. Uno puede desplazarse hacia arriba o hacia abajo por la lista del cuadro para hacer desplegar la sintaxis y una breve descripción de cada una de las funciones estadísticas. Por ejemplo, al desplazarse hacia abajo se puede seleccionar la función COUNTIF como se muestra en la figura 2. Observe que ahora aparece sombreada la función COUNTIF y que inmediatamente abajo del cuadro aparece **COUNTIF**(rango y criterio), lo que indica que la función COUNTIF contiene dos argumentos, rango y criterio. Además se ve que la descripción de la función COUNTIF es "Count the number of cells within a range that meet the given condition" ("Cuenta la cantidad de celdas dentro de un determinado rango que satisfacen la condición dada").

En las versiones anteriores de Excel aparece un cuadro de diálogo parecido. Ese cuadro tiene el mismo propósito que el cuadro de diálogo ***Function Arguments*** *de Excel 2003.*

Si la función seleccionada (sombreada) es la que se desea usar, se hace clic en **OK**; entonces aparece el cuadro de diálogo **Argumentos de la función**. En la figura 3 se muestra el cuadro de diálogo Argumentos de la función correspondiente a la función COUNTIF. Este cuadro de diálogo ayuda a dar los argumentos correspondientes a la función seleccionada. Al terminar de ingresar los argumentos, se hace clic en **OK**; entonces Excel inserta la función en la celda de la hoja de cálculo.

Insertar una función en una celda de la hoja de cálculo

Se mostrará cómo usar los cuadros de diálogo Insertar función y Argumentos de la función para seleccionar una función, dar sus argumentos e insertar la función en una celda de una hoja de cálculo.

En la sección 2.1, se usó la función COUNTIF de Excel para elaborar una distribución de frecuencias para las compras de refrescos. En la figura 4 se presenta una hoja de cálculo de Excel con los datos de refrescos y los rótulos para la distribución de frecuencias que se quiere cons-

FIGURA 2 DESCRIPCIÓN DE LA FUNCIÓN COUNTIF EN LA CAJA DE DIÁLOGO INSERT FUNCTION

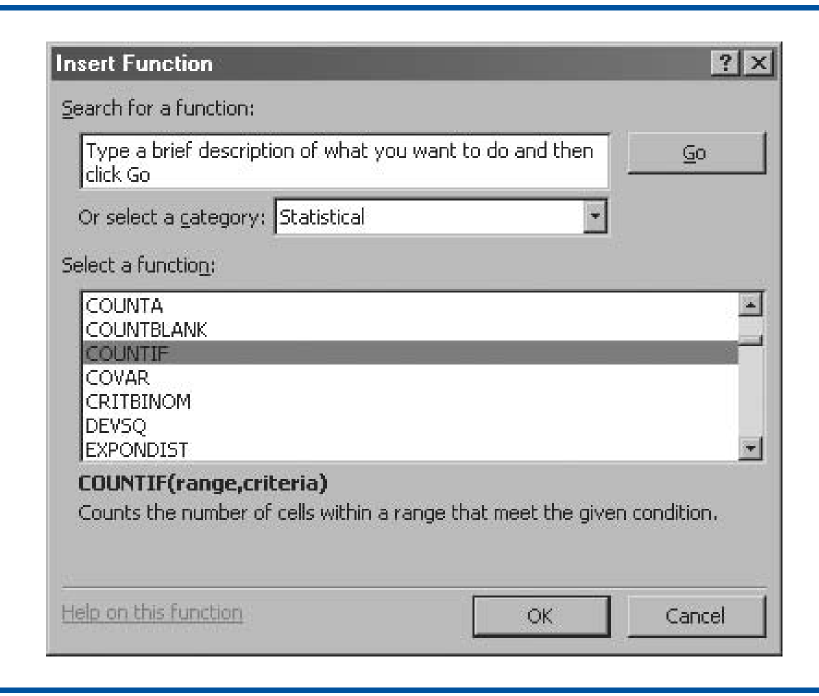


FIGURA 3 CUADRO DE DIÁLOGO PARA LOS ARGUMENTOS DE LA FUNCIÓN COUNTIF

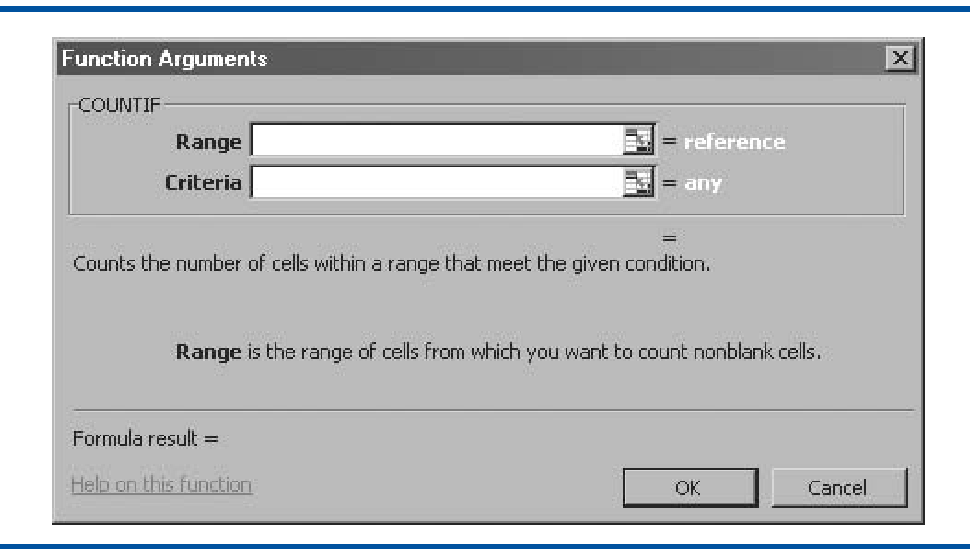


FIGURA 4 HOJA DE CÁLCULO DE EXCEL CON LOS DATOS DE LOS REFRESCOS Y LOS RÓTULOS PARA LA DISTRIBUCIÓN DE FRECUENCIAS QUE SE DESEA CONSTRUIR



Nota: Los renglones 11 a 44 están ocultos.

	A	B	C	D	E
1	**Brand Purchased**		**Soft Drink**	**Frequency**	
2	Coke Classic		Coke Classic		
3	Diet Coke		Diet Coke		
4	Pepsi		Dr. Pepper		
5	Diet Coke		Pepsi		
6	Coke Classic		Sprite		
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

truir. Como se ve, la frecuencia de las compras de Coke Clasic irán en la celda D2, la frecuencia de Diet Coke irán en la celda D3, y así sucesivamente. Suponga que se quiere usar la función COUNTIF para calcular las frecuencias de estas celdas y que se desea tener ayuda de Excel.

Paso 1. Seleccionar la celda D2
Paso 2. Clic en f_x en la barra de la herramientas (o seleccionar **Insertar** y después elegir **Función**)
Paso 3. Cuando aparezca el cuadro de diálogo **Insertar función**:
Seleccionar **Estadísticas** en el cuadro **O seleccionar categoría**
Seleccionar **COUNTIF** en el cuadro **Seleccionar una función**
Clic en **OK**
Paso 4. Cuando aparezca el cuadro **Argumentos de la función** (ver figura 5):
Ingresar $A$2:$A$51 en el cuadro **Rango**
Ingresar C2 en el cuadro **Criterio** (En este momento, en el penúltimo renglón del cuadro de diálogo aparecerá el valor de la función. Su valor es 19.)
Clic en **OK**
Paso 5. Copiar las celda D2 a las celdas D3:D6

Entonces la hoja de cálculo se verá como en la figura 6. La hoja de cálculo con las fórmulas aparece al fondo y la hoja de cálculo con los valores aparece al frente. En la hoja de cálculo con las fórmulas se ve que la función COUNTIF ha sido insertada en la celda D2. El contenido de la celda D2 se ha copiado a las celdas D3:D6. En la hoja de cálculo con los valores aparecen las frecuencias de las clases calculadas.

Se ha ilustrado el uso de Excel para obtener ayuda al usar la función COUNTIF. Con todas las demás funciones de Excel el procedimiento es similar. Esta posibilidad es especialmente útil cuando no se sabe qué función usar o cuando se ha olvidado el nombre o la sintaxis de una función.

FIGURA 5 CUADRO DE DIÁLOGO PARA DAR LOS ARGUMENTOS DE LA FUNCIÓN COUNTIF

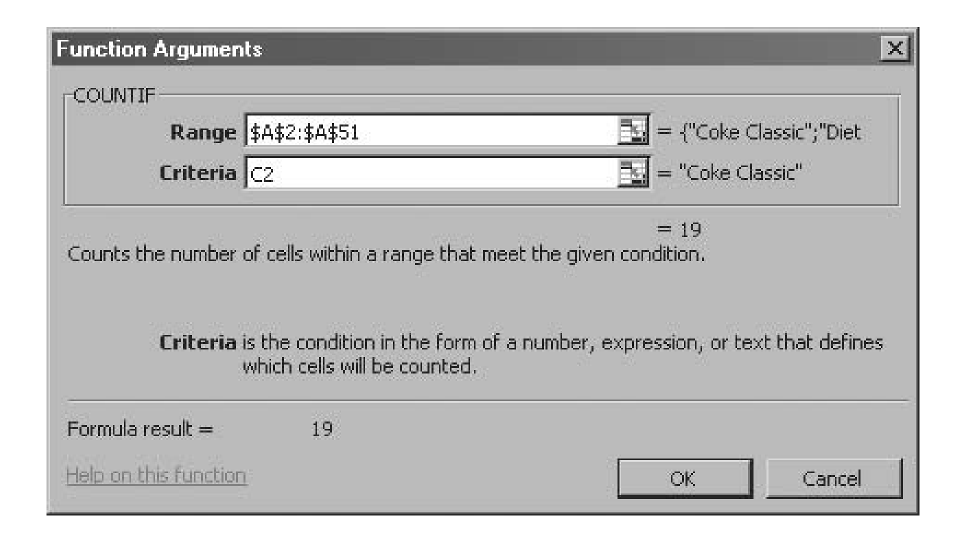


FIGURA 6 HOJA DE CÁLCULO DE EXCEL EN LA QUE SE MUESTRA EL USO DE LA FUNCIÓN COUNTIF DE EXCEL PARA ELABORAR UNA DISTRIBUCIÓN DE FRECUENCIAS

	A	B	C	D	E
1	**Brand Purchased**		**Soft Drink**	**Frequency**	
2	Coke Classic		Coke Classic	=COUNTIF($A$2:$A$51,C2)	
3	Diet Coke		Diet Coke	=COUNTIF($A$2:$A$51,C3)	
4	Pepsi		Dr. Pepper	=COUNTIF($A$2:$A$51,C4)	
5	Diet Coke		Pepsi	=COUNTIF($A$2:$A$51,C5)	
6	Coke Classic		Sprite	=COUNTIF($A$2:$A$51,C6)	
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

Nota: Los renglones 11 a 44 están ocultos.

	A	B	C	D	E
1	**Brand Purchased**		**Soft Drink**	**Frequency**	
2	Coke Classic		Coke Classic	19	
3	Diet Coke		Diet Coke	8	
4	Pepsi		Dr. Pepper	5	
5	Diet Coke		Pepsi	13	
6	Coke Classic		Sprite	5	
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

Apéndice F: Cálculo de los valores-p usando Minitab o Excel

Aquí se describe cómo usar Minitab o Excel para calcular los valores-p correspondientes a los estadísticos z, t, χ^2 y F que se usan en las pruebas de hipótesis. Como se dijo en el libro, con las tablas sólo es posible obtener valores-p aproximados correspondientes a los estadísticos t, χ^2 y F. Este apéndice es útil para las personas que han calculado los estadísticos de prueba a mano, o por otros medios, y que desean emplear un software para obtener el valor-p exacto.

Uso de Minitab

Minitab puede usarse para obtener la probabilidad acumulada correspondiente a los estadísticos z, t, χ^2 y F. El valor-p en la cola inferior se obtiene directamente. El valor-P en la cola superior se calcula restando de 1 el valor-p encontrado para la cola inferior. El valor-p en las dos colas se obtiene multiplicando, por dos, el menor de los valores-p de las colas superior e inferior.

Estadístico de prueba z. Como ilustración se empleará la prueba de hipótesis de la cola inferior de Hilltop Coffee, de la sección 9.3; el valor del estadístico de prueba es $z = -2.67$. Los pasos a seguir para calcular la probabilidad acumulada correspondiente a $z = -2.67$ son los siguientes.

Paso 1. Seleccionar el menú **Calc**
Paso 2. Elegir **Probability distributions**
Paso 3. Elegir **Normal**
Paso 4. Cuando aparezca el cuadro de diálogo Normal Distribution:
Seleccionar **Cumulative probability**
Ingresar 0 en el cuadro **Mean**
Ingresar 1 en el cuadro **Standard deviation**
Seleccionar **Input Constant**
Ingresar –2.67 en el cuadro **Input Constant**
Hacer clic en **OK**

Minitab dará como probabilidad acumulada 0.0038. Esta probabilidad acumulada es el valor-p para la cola inferior que se usó en la prueba de hipótesis de Hilltop Coffee.

Para una prueba de la cola superior, el valor-p se obtiene como sigue, a partir de la probabilidad acumulada obtenida de Minitab.

$$\text{Valor-}p = 1 - \text{probabilidad acumulada}$$

Por ejemplo, el valor-p en la cola superior, correspondiente al estadístico de prueba $z = -2.67$ es $1 - 0.0038 = 0.9962$. El valor-p para dos colas correspondiente al estadístico de prueba $z = -2.67$ es el doble del mínimo de los valores-p correspondientes a las colas inferior y superior; es decir, el valor-p para dos colas correspondiente a $z = -2.67$ es $2(0.0038) = 0.0076$.

Estadístico de prueba t. Como ilustración se usará el ejemplo del aeropuerto de Heathrow, de la sección 9.4; el valor del estadístico de prueba es $t = 1.84$ con 59 grados de libertad. Para emplear Minitab para calcular la probabilidad acumulada correspondiente a $t = 1.84$ se siguen los pasos que se presentan a continuación.

Paso 1. Seleccionar el menú **Calc**
Paso 2. Elegir **Probability Distributions**

Paso 3. Elegir ***t***
Paso 4. Cuando aparezca el cuadro de diálogo t Distribution:
Seleccionar **Cumulative probability**
Ingresar 59 en el cuadro **Degrees of freedom**
Seleccionar **Input Constant**
Ingresar 1.84 en el cuadro **Input Constant**
Clic en **OK**

Minitab da la probabilidad acumulada, que es 0.9646, y por tanto el valor-*p* en la cola inferior es 0.9646. En el ejemplo del aeropuerto de Heatrow se trata de una prueba de la cola superior. El valor-*p* para la cola superior es 1 − 0.9646 = 0.0354. En el caso de una prueba de dos colas se usará el mínimo de 0.9646 y 0.0354 para obtener el valor-*p* que es 2(0.0354) = 0.0708.

Estadístico de prueba χ^2. Como ilustración se empleará el ejemplo del metrobús de St. Louis, de la sección 11.1; el valor del estadístico de prueba es $\chi^2 = 28.18$ con 23 grados de libertad. Para calcular la probabilidad acumulada correspondiente a $\chi^2 = 28.18$ empleando Minitab se siguen los pasos que se dan a continuación.

Paso 1. Seleccionar el menú **Calc**
Paso 5. Elegir **Probability Distributions**
Paso 2. Elegir **Chi-Square**
Paso 3. Cuando aparezca el cuadro de diálogo Chi-Square Distribution:
Seleccionar **Cumulative probability**
Ingresar 23 en el cuadro **Degrees of freedom**
Seleccionar **Input Constant**
Ingresar 28.18 en el cuadro **Input Constant**
Clic en **OK**

Minitab da la probabilidad acumulada, que es 0.7909, que es el valor-*p* correspondiente a la cola inferior. El valor-*p* en la cola superior es 1 − probabilidad acumulada o 1 − 0.7909 = 0.2091. El valor-*p* para una prueba de dos colas es 2 multiplicado por el mínimo del valor-*p* en las colas inferior y superior. Por tanto, el valor-*p* para dos colas es 2(0.2091) = 0.4182. En el ejemplo del metrobús de St. Louis se tiene una prueba de la cola superior, por lo que el valor-*p* es 0.2091

El estadístico de prueba *F*. Como ilustración se usará el ejemplo del las escuelas Dullus County, de la sección 11.2; el valor del estadístico de prueba es $F = 2.40$ con 25 grados de libertad en el numerador y 15 grados de libertad en el denominador. Para calcular la probabilidad acumulada correspondiente a $F = 2.40$ empleando Minitab se siguen los pasos que se dan a continuación.

Paso 1. Seleccionar el menú **Calc**
Paso 2. Elegir **Probability Distributions**
Paso 3. Elegir ***F***
Paso 4. Cuando aparezca el cuadro de diálogo F Distribution:
Seleccionar **Cumulative probability**
Ingresar 25 en el cuadro **Numerator degrees of freedom**
Ingresar 15 en el cuadro **Denominator degrees of freedom**
Seleccionar **Input Constant**
Ingresar 2.40 en el cuadro **Input Constant**
Clic en **OK**

Minitab da la probabilidad acumulada y, por tanto, el valor-*p* en la cola inferior, que es 0.9594. El valor-*p* en la cola superior es 1 − 0.9594 = 0.0406. Como en el ejemplo de las escuelas Dullus County se trata de una prueba de dos colas se usa el mínimo de 0.9594 y 0.0406 para obtener el valor-*p* que es 2(0.0406) = 0.0812.

Uso de Excel



p-Value

Las funciones y las fórmulas de Excel pueden emplearse para calcular los valores-*p* correspondientes a los estadísticos de prueba z, t, c^2 y F. En el archivo del disco compacto titulado valores-*p* se proporciona una plantilla para calcular los valores-*p*. Con el uso de esta plantilla sólo es necesario ingresar el valor del estadístico de prueba y, en caso necesario, los grados de libertad. Consulte la figura F.1 a medida que lee la descripción de cómo usar la plantilla. Aquellos usuarios interesados en las funciones y fórmulas de Excel empleadas, sólo necesitan hacer clic sobre la celda correspondiente de la plantilla.

Estadístico de prueba z. Como ilustración se usará la prueba de hipótesis de la cola inferior de Hilltop Coffee, de la sección 9.3; el valor del estadístico de prueba es $z = -2.67$. Para hacer uso de la plantilla, simplemente ingrese −2.67 en la celda B6 (véase figura F.1). Una vez hecho esto, aparecerán los valores-*p* para los tres tipos de pruebas de hipótesis. Para el problema de Hilltop Coffee, se usará el valor-*p* de la cola inferior, que es 0.0038 y que aparece en la celda B9. Para una prueba de la cola superior se usará el valor-*p* que aparece en la celda B10, y para una prueba de dos colas se usará el valor-*p* que aparece en la celda B11.

Estadístico de prueba t. Como ilustración se usará el ejemplo del aeropuerto de Heathrow, de la sección 9.4; el valor del estadístico de prueba es $t = 1.84$ con 59 grados de libertad. Para hacer uso de la plantilla, simplemente ingrese 1.84 en la celda E6 y 59 en la celda E7 (véase figura F.1). Una vez hecho esto, aparecerán los valores-*p* para los tres tipos de pruebas de hipótesis. En

FIGURA F.1 HOJA DE TRABAJO DE EXCEL PARA CALCULAR VALORES-*p*

	A	B	C	D	E
1	**Computing *p*-Values**				
2					
3					
4	**Using the Test Statistic *z***			**Using the Test Statistic *t***	
5					
6	**Enter *z* -->**	−2.67		**Enter *t* -->**	1.84
7				***df* -->**	59
8					
9	***p*-value (Lower Tail)**	0.0038		***p*-value (Lower Tail)**	0.9646
10	***p*-value (Upper Tail)**	0.9962		***p*-value (Upper Tail)**	0.0354
11	***p*-value (Two Tail)**	0.0076		***p*-value (Two Tail)**	0.0708
12					
13					
14					
15					
16	**Using the Test Statistic Chi Square**			**Using the Test Statistic *F***	
17					
18	**Enter Chi Square -->**	28.18		**Enter *F* -->**	2.40
19	**df -->**	23		**Numerator df -->**	25
20				**Denominator df -->**	15
21					
22	***p*-value (Lower Tail)**	0.7909		***p*-value (Lower Tail)**	0.9594
23	***p*-value (Upper Tail)**	0.2091		***p*-value (Upper Tail)**	0.0406
24	***p*-value (Two Tail)**	0.4181		***p*-value (Two Tail)**	0.0812

el ejemplo del aeropuerto de Heathrow se trata de una prueba de la cola superior, el valor-*p* es 0.0354 y aparece en la celda E10.

Estadístico de prueba χ^2. Como ilustración se empleará el ejemplo del metrobús de St. Louis, de la sección 11.1; el valor del estadístico de prueba es $c^2 = 28.18$ con 23 grados de libertad. Para hacer uso de la plantilla, simplemente ingrese 28.18 en la celda B18 y 23 en la celda B19 (véase figura F.1). Una vez hecho esto, aparecerán los valores-*p* para los tres tipos de pruebas de hipótesis. En el ejemplo del metrobús de St. Louis se trata de una prueba de la cola superior, por lo que se usará el valor-$p = 0.2091$ que aparece en la celda B23.

Estadístico de prueba *F*. Como ilustración se usará el ejemplo de las escuelas Dullus County, de la sección 11.2; el valor del estadístico de prueba es $F = 2.40$ con 25 grados de libertad en el numerador y 15 grados de libertad en el denominador. Para hacer uso de la plantilla, simplemente ingrese 2.40 en la celda E18, ingrese 25 en la celda E19 e ingrese 15 en la celda B20 (véase figura F.1). Una vez hecho esto, aparecerán los valores-*p* para los tres tipos de pruebas de hipótesis. En el ejemplo de las escuelas Dullus County se trata de una prueba de dos colas, de manera que se usará el valor-*p* para dos colas, que es 0.0812 y aparece en la celda E24.

Índice

Números de página seguidos de una **n** indican que se trata de una nota a pie de página.

A

B

C

D

E

F

G

H

I

N

O

P

R

S

T

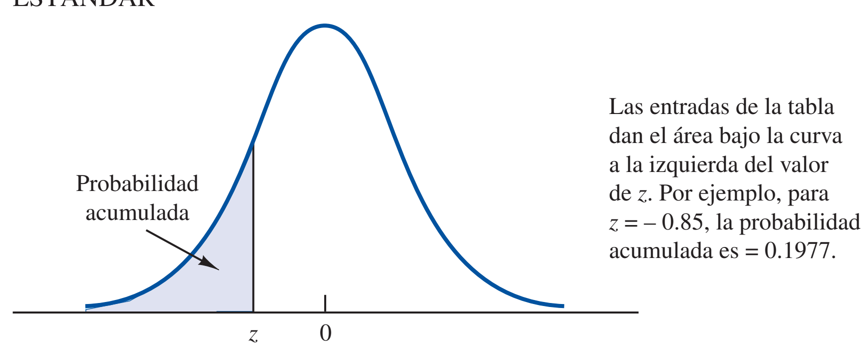
U

V

W

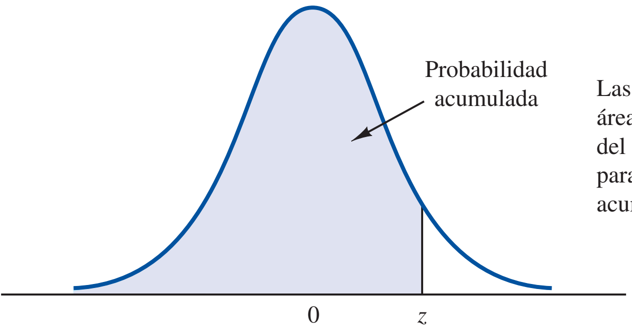
Z

PROBABILIDADES ACUMULADAS DE LA DISTRIBUCIÓN NORMAL ESTÁNDAR



z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

PROBABILIDADES ACUMULADAS DE LA DISTRIBUCIÓN NORMAL ESTÁNDAR



Las entradas de la tabla dan el área bajo la curva a la izquierda del valor de z. Por ejemplo, para $z = 1.25$, la probabilidad acumulada es = 0.8944.

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9913
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9986	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Disco compacto que acompaña a *Estadística para administración y economía*

Capítulo 1

BWS&P	Tabla 1.1
Hotel	Tabla 1.6
Minisystems	Tabla 1.7
Norris	Tabla 1.5
Shadow02	Ejercicio 25

Capítulo 2

ApTest	Tabla 2.8
Audit	Tabla 2.4
AutoData	Ejercicio 38
Broker	Ejercicio 26
CEOs	Ejercicio 9
CityTemp	Ejercicio 46
Computer	Ejercicio 21
Concerts	Ejercicio 20
Crosstab	Ejercicio 29
DivYield	Ejercicio 41
Fortune	Ejercicio 51
Frequency	Ejercicio 11
Golf	Ejercicio 40
Holiday	Ejercicio 18
IBD	Ejercicio 34
Major	Ejercicio 39
Marathon	Ejercicio 28
Movies	Caso problema 2
Names	Ejercicio 5
Networks	Ejercicio 6
NFL	Ejercicio 37
OccupSat	Ejercicio 48
PelicanStores	Caso problema 1
Population	Ejercicios 44
PriceShare	Ejercicio 17
Restaurant	Tabla 2.9
RevEmps	Ejercicio 49
SATScores	Ejercicio 42
Scatter	Ejercicio 30
Shadow	Ejercicio 43
SoftDrink	Tabla 2.1
Stereo	Tabla 2.12
TVMedia	Ejercicio 4

Capítulo 3

Ages	Exersice 59
Asian	Caso problema 3
BASalary	Ejercicio 6
Baseball	Ejercicio 42
Beer	Ejercicio 65
Broker	Ejercicios 7 & 22
Disney	Ejercicio 12
Homes	Ejercicio 64
Hotels	Ejercicio 5
Movies	Caso problema 2
Mutual	Ejercicio 44
NCAA	Ejercicio 34
PCs	Ejercicio 49
PelicanStores	Caso problema 1
Penalty	Ejercicio 62
Property	Ejercicio 40
Speakers	Ejercicio 35
StartSalary	Tabla 3.1
Stereo	Tabla 3.7
StockMarket	Ejercicio 50
Temperature	Ejercicio 51
Visa	Ejercicio 58

Capítulo 4

Judge	Caso problema

Capítulo 6

Volume	Ejercicio 24

Capítulo 7

EAI	Sección 7.1
MetAreas	Tabla 7.6
MutualFund	Ejercicio 14

Capítulo 8

ActTemps	Ejercicio 49
Alcohol	Ejercicio 21
Auto	Caso problema 3
FastFood	Ejercicio 18
Flights	Ejercicio 48
GulfProp	Caso problema 2
Interval p	Apéndice 8.2
JobSatisfaction	Ejercicio 37
Lloyd's	Sección 8.1
Miami	Ejercicio 17
NewBalance	Tabla 8.3
Nielsen	Ejercicio 6
NYSEStocks	Ejercicio 47
OpenEndFunds	Ejercicio 22
Professional	Caso problema 1
Program	Ejercicio 20
Scheer	Tabla 8.4
TeeTimes	Sección 8.4

Capítulo 9

AirRating	Sección 9.4
BLS	Caso problema 2
Coffee	Sección 9.3
Diamonds	Ejercicio 29
Drowsy	Ejercicio 44
Eagle	Ejercicio 43
Fowle	Ejercicio 21
Gasoline	Ejercicio 67
GolfTest	Sección 9.3
Hyp Sigma Known	Apéndice 9.2
Hyp Sigma Unknown	Apéndice 9.2
Hypothesis p	Apéndice 9.2
Orders	Sección 9.4
Quality	Caso problema 1
RentalRates	Ejercicio 16
UsedCars	Ejercicio 32
WomenGolf	Sección 9.5

Capítulo 10

AirFare	Ejercicio 24
Cargo	Ejercicio 13
CheckAcct	Sección 10.2
Digital	Ejercicio 39
Earnings	Ejercicio 26
Earnings2005	Ejercicio 22
ExamScores	Sección 10.1
Florida	Ejercicio 42
Golf	Caso problema
Matched	Tabla 10.2
Mortgage	Ejercicio 6
Mutual	Ejercicio 40
SAT	Ejercicio 18
SATVerbal	Ejercicio 16
SoftwareTest	Tabla 10.1
TaxPrep	Sección 10.4
TVRadio	Ejercicio 25

Capítulo 11

Bags	Ejercicio 19
BusTimes	Sección 11.1
Return	Ejercicio 6
SchoolBus	Sección 11.2
Training	Caso problema
Travel	Ejercicio 25

Capítulo 12

Chemline	Tabla 12.10
FitTest	Apéndice 12.2
Independence	Apéndice 12.2
NYReform	Caso problema

Capítulo 13

AirTraf	Tabla 13.5
Assembly	Ejercicio 38
AudJudg	Ejercicio 10
Browsing	Ejercicio 39
Chemitech	Tabla 13.1
Exer6	Ejercicio 6
Funds	Ejercicio 36
GMAT	Tabla 13.10
Grocery	Ejercicio 41
JobSalary	Ejercicio 37
Medical1	Caso problema 1
Medical2	Caso problema 1
NCP	Tabla 13.4
Paint	Ejercicio 11
Salaries	Ejercicio 32
SalesSalary	Caso problema 2
SatisJob	Ejercicio 35
Ships	Ejercicio 20
Snow	Ejercicio 27
Vitamins	Ejercicio 25

Capítulo 14

Absent	Ejercicio 63
ADRs	Ejercicio 49
AgeCost	Ejercicio 64
Airport	Ejercicio 11
Alumni	Caso problema 3
Armand's	Tabla 14.1
Beer	Ejercicio 52
Beta	Caso problema 1
Boats	Ejercicio 5
Boots	Ejercicio 27
Cars	Ejercicios 7 & 19
CEO	Ejercicio 54
DJIAS&P500	Ejercicio 60
HoursPts	Ejercicio 65
Hydration1	Ejercicio 43
Hydration2	Ejercicio 53
IPO	Ejercicio 58
IRSAudit	Ejercicio 67
Jensen	Ejercicio 61
JetSki	Ejercicio 12
JobSat	Ejercicio 68
MktBeta	Ejercicio 66
MLB	Caso problema 4
MtnBikes	Ejercicio 8
OffRates	Ejercicio 44
Options	Ejercicio 59
PlasmaTV	Ejercicios 20 & 31
Printers	Ejercicios 22 & 30
Safety	Caso problema 2
Salaries	Ejercicio 14
Sales	Ejercicio 9
SleepingBags	Ejercicios 10, 28 & 36
VPSalary	Ejercicio 6

Capítulo 15

Alumni	Caso problema 3
Auto2	Ejercicio 42
Backpack	Ejercicio 7
Bank	Ejercicio 46
Boats	Ejercicios 9, 17 & 30
Brokers	Ejercicio 25
Butler	Tablas 15.1 & 15.2
Chocolate	Ejercicio 48
Consumer	Caso problema 1
Enquirer	Caso problema 2
Exer2	Ejercicio 2
Football	Ejercicio 37
ForFunds	Ejercicio 8
FuelEcon	Ejercicio 56
HomeValue	Ejercicio 54
Johnson	Tabla 15.6
Lakeland	Ejercicio 47
LPGA	Ejercicio 43
MLB	Ejercicios 6, 16 & 24
NBA	Ejercicios 10, 18 & 26
NFLStats	Caso problema 4
Repair	Ejercicio 35
Showtime	Ejercicios 5, 15 & 41
Simmons	Tabla 15.11
SportsCar	Ejercicio 31
Stroke	Ejercicio 38
Treadmills	Ejercicio 55
Trucks	Ejercicio 57

Capítulo 16

Audit	Ejercicio 31
Browsing	Ejercicio 34
Cars	Caso problema 2
ClassicCars	Ejercicio 8
ColorPrinter	Ejercicio 29
Cravens	Tabla 16.5
GradRate	Caso problema 3
IBM	Ejercicio 27
Layoffs	Ejercicio 16
LightRail	Ejercicio 9
LPGATour	Ejercicios 12 & 13
LPGATour2	Ejercicio 17
Monitors	Ejercicio 7
MPG	Tabla 16.4
MutFunds	Ejercicio 30
NFL	Ejercicio 15
PGATour	Caso problema 1
Resale	Ejercicio 35
Reynolds	Tabla 16.1
Stroke	Ejercicios 14 & 19
Tyler	Tabla 16.2
Yankees	Ejercicio 18

Capítulo 18

AcctBal	Ejercicio 35
AptExp	Ejercicio 24
Bicycle	Tabla 18.6
Broadband	Ejercicio 18
CDSales	Ejercicio 32
Gasoline	Tabla 18.1
IBM	Ejercicio 27
MfgCap	Ejercicio 11
Pollution	Ejercicio 25
Power	Ejercicio 26
TVSales	Tabla 18.7
Vintage	Caso problema 1

Capítulo 19

Annual	Ejercicio 11

Capítulo 20

Coffee	Ejercicio 20
Jensen	Tabla 20.2
Tires	Ejercicio 7

Capítulo 21

PDC Tree	Apéndice 21.1

Apéndice F

p-Value	Apéndice F

La nueva edición de esta obra, un verdadero *best-seller*, tanto en Estados Unidos como en América Latina, continúa presentando una gran cantidad de ejercicios con datos reales actualizados. Las secciones de problemas se dividen en tres partes a fin de reforzar lo aprendido: métodos, aplicaciones y autoevaluaciones. Además contiene secciones y advertencias sobre los errores estadísticos más comunes en los que se puede incurrir.

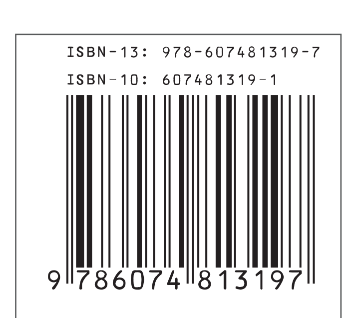
Características

- A lo largo de todo el texto se plantean situaciones de negocios y económicas reales.
- Se muestra el uso de la computadora; especialmente se enfatiza el trabajo con Excel y con MINITAB en sus versiones más recientes.
- Presenta una mayor cobertura en métodos tabulares y gráficos de la estadística descriptiva.
- Integra el uso de Excel para el muestreo aleatorio.
- Incorpora el uso de apoyos en línea integrados a lo largo del texto.
- Un nuevo apéndice F cubre el uso de software para calcular el valor de p y muestra claramente el uso de MINITAB y Excel para calcular los valores de p asociados a pruebas estadísticas z, t y F.
- Emplea software estadístico para el uso de tablas de distribución normal acumulada, lo que hace más sencillo para el alumno el cálculo de los valores de p en las pruebas de hipótesis.
- Integra casos al final de cada capítulo.

Éste es sin duda el mejor libro de Estadística para Administración y Economía en español.



http://latinoamerica.cengage.com


ISBN-13: 978-607481319-7
ISBN-10: 607481319-1